DOCUMENT RESUME

ED 101 649 HE 006 250

AUTHOR Sockloff, Alan L., Ed.
TITLE Proceedings: Faculty Effectiveness as Evaluated by Students.
INSTITUTION Temple Univ., Philadelphia, Pa. Measurement and Research Center.
PUB DATE [73]
NOTE 259p.; Proceedings of the First Invitational Conference on Faculty Effectiveness as Evaluated by Students (Temple University, Philadelphia, Pennsylvania, April 1973)

EDRS PRICE MF-$0.76 HC-$13.32 PLUS POSTAGE
DESCRIPTORS Affective Behavior; College Students; Conference Reports; Educational Improvement; *Effective Teaching; *Faculty; *Faculty Evaluation; Feedback; *Higher Education; Instructional Improvement; Models; Performance Criteria; Student Attitudes; Student Opinion; Teacher Evaluation; Teacher Rating; *Teaching Quality
IDENTIFIERS *Kansas State University; Michigan State University

ABSTRACT

Faculty effectiveness as evaluated by students was the focal point of the first invitational conference sponsored by the Measurement and Research Center of Temple University. Papers presented cover: the rationale of student evaluation of faculty, the impact of student ratings on academia, the usefulness of student evaluations in improving college teaching, some considerations and a model of faculty evaluation, a system for helping teachers to change their affective behavior through feedback, instruments for student evaluation of faculty, the Kansas State University Program for assessing and improving instructional effectiveness, student evaluation of instruction at Michigan State University, criteria for evaluation of college teaching, correlates of student ratings, the shortcomings of traditional approaches to faculty evaluation and faculty performance under stress. (MJM)

ED101649

The Measurement and Research Center of Temple University

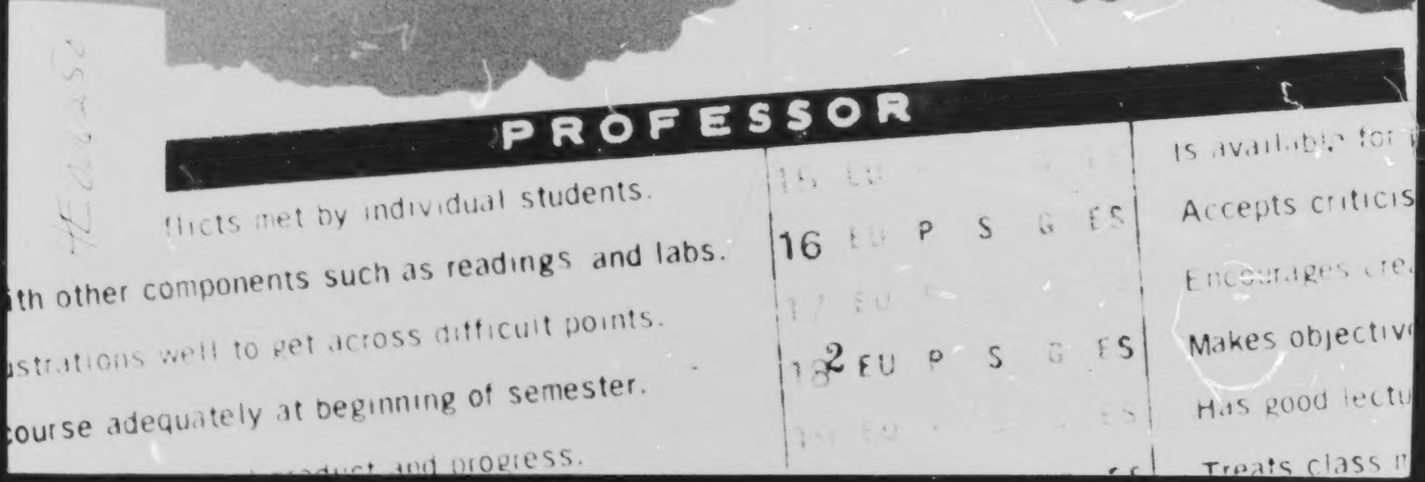
PROCEEDINGS

Alan L. Sockloff, Editor

The First Invitational Conference on FACULTY EFFECTIVENESS AS EVALUATED BY STUDENTS

U.S. DEPARTMENT OF HEALTH,
EDUCATION & WELFARE
NATIONAL INSTITUTE OF
EDUCATION
THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL NATIONAL INSTITUTE OF EDUCATION POSITION OR POLICY.

Temple University



PROFESSOR

flicts met by individual students.

ith other components such as readings and labs.

strations well to get across difficult points.

course adequately at beginning of semester.

product and progress.

15 EU

16 EU P S G ES

17 EU

18 EU P S G ES

2

19 EU

Is available for

Accepts criticis

Encourages cre

Makes objective

Has good lectu

Treats class

© by the Measurement and Research Center
Temple University

3/4

CONTENTS

Preface v

I. General

Student Evaluation of Faculty: Why? What? How?
PAUL L. DRESSEL 1

II. Impact

The Student as Godfather? The Impact of Student Ratings on Academia
JOHN A. CENTRA 25

The Usefulness of Student Evaluations in Improving College Teaching
LAWRENCE M. ALEAMONI 42

III. Systems

Faculty Evaluation: Some Considerations and a Model
KENNETH O. DOYLE, JR. 59

A System for Helping Teachers to Change their Affective Behavior through Feedback
BRUCE W. TUCKMAN 91

IV. Instruments

Instruments for Student Evaluation of Faculty: Ideal and Actual
ALAN L. SOCKLOFF 132

The Kansas State University Program for Assessing and Improving Instructional Effectiveness
DONALD P. HOYT 152

Student Evaluation of Instruction at Michigan State University
WILLARD G. WARRINGTON 164

Criteria for Evaluation of College Teaching: Their Reliability and Validity at the University of Toledo
RICHARD R. PERRY and REEMT R. BAUMANN 183

V. Correlates

Correlates of Student Ratings
WILBERT J. MCKEACHIE 213

Teachers Who Make a Difference
JERRY G. GAFF 219

Faculty Performance Under Stress
MARY JO CLARK AND ROBERT T. BLACKBURN 233

CONTRIBUTING AUTHORS

LAWRENCE M. ALEAMONI, Head, Measurement and Research Division and Professor of Educational Psychology, University of Illinois

REEMT R. BAUMANN, Professor and Chairman, Department of Educational Research and Measurement, University of Toledo

ROBERT T. BLACKBURN, Professor of Higher Education, University of Michigan

JOHN A. CENTRA, Research Psychologist, Educational Testing Service

MARY JO CLARK, Associate Research Psychologist, Educational Testing Service

KENNETH O. DOYLE, JR., Research Associate, Measurement Services Center, University of Minnesota

PAUL L. DRESSEL, Director, Office of Institutional Research and Assistant Provost, Michigan State University

JERRY G. GAFF, Visiting Professor, Department of Nursing, Sonoma State College

DONALD P. HOYT, Director, Office of Educational Resources, Kansas State University

WILBERT J. McKEACHIE, Professor of Psychology, University of Michigan

RICHARD R. PERRY, Assistant Vice President for Academic Affairs and Professor of Higher Education, University of Toledo

ALAN L. SOCKLOFF, Director of Psychometric Research, Measurement and Research Center and Assistant Professor of Psychology, Temple University

BRUCE W. TUCKMAN, Professor of Vocational Education, Rutgers University

WILLARD G. WARRINGTON, Director, Office of Evaluation Services and Assistant Dean, University College, Michigan State University

PREFACE

Between April 25 and 27, 1973, the Measurement and Research Center of Temple University held its First Invitational Conference "Faculty Effectiveness as Evaluated by Students." As is obvious from the title of the Conference, the major interest was in the student component of faculty evaluation. Basically, there were two reasons for this "narrow" approach. First, there exists a more extensive body of knowledge regarding evaluation of faculty by students. By default, information on administrative and colleague ratings is limited and is rarely within the public domain. Second, assuming that the issue of faculty evaluation cannot be comprehensively covered or even partially resolved in a single conference, it was considered preferable to concentrate on this small component about which something is known rather than become involved in the kinds of philosophical speculation that typically adhere to the broader issues.

In planning the Conference, the decision was made to concentrate upon five broad areas, and speakers were gathered to cover these broad areas at an overview level. Recognizing that treatment at an overview level cannot be expected to produce anything more than overview knowledge, a search was begun to find additional speakers who could present papers that were more specific and more practical. In deciding upon these additional speakers, the guiding considerations were toward programmatic series of research studies, as well as quality and uniqueness of approach.

With a minimum of persuasion, all speakers had complete flexibility in their presentations. However, the initial delineation of the subject into five broad areas may have been unrealistic insofar as researchers in faculty evaluation rarely focus their efforts in areas as specific as those outlined for the Conference. Even so, it was hoped that, given a certain amount of overlap, the presentations would be sufficiently distinct to warrant their placement within the five areas.

The results of the Conference were mildly surprising. The distinctness of the five areas appeared to be at a minimum, and this may have been partly due to both the state of present knowledge of student evaluation of faculty and the impossible desire to delineate the presentations into five distinct, but broad, areas. Examination of the 12 papers presented in this volume suggests common themes; but, if compared against what one would expect from a conference designed to partially resolve at least some of the issues, the extent of agreement among papers was relatively small. Related to this difficulty, analysis of the taped transcriptions of the discussion following the presentations revealed an almost meaningless sequence of verbalizations. Perhaps, the mix of the participants, "lay" researchers, administrators, and psychometric types, compounded this problem.

The desired order of the sessions was: General Keynote, Impact, Systems, Instruments, Correlates, and Discussion of Issues. However, timing problems arose, and the following order was used: General Keynote, Impact, Instruments, Correlates, Discussion of Issues, and Systems. The papers contained within this volume follow the originally desired order, but with one minor change. Donald Hoyt's paper was presented under Systems, but is contained in this volume under Instruments because its contents are more closely aligned with problems of instrument construction.

The Conference was somewhat pessimistically opened with Paul Dressel's Keynote talk. Playing the devil's advocate, Dressel pointed out the breadth of faculty responsibility and the small role played by classroom teaching. Raising more questions than could be answered within the context of the Conference, one of Dressel's criticisms is that faculty evaluation, as presently done, is dominated by an inadequate conception of teaching and is not very useful for the improvement of teaching. Responses to Dressel's plaints come from the papers of John Centra and Lawrence Aleamoni in the Impact session. If Dressel is correct, then the impact of faculty evaluation in the domain of teaching must be nil. Centra's presentation and discussion of the many facets of impact stands as a more optimistic outlook, while Aleamoni's presentation shows evidence of early research into this foundling area.

The rationale behind the session on Systems was to inquire into how faculty evaluation is, and can be, done from a molar systems perspective, over and above the technical details of instrument construction. Kenneth Doyle's paper presents a comprehensive description of many of the considerations that should go into creating a faculty evaluation system, while Bruce Tuckman's paper evidences the logic and approach to developing a micro-system for changing affective responses of faculty. A unique aspect of Tuckman's paper is its attempt to develop a system based on theory from another field. We can see that Systems is a relatively new, and relatively unresearched, area begging for more substance in order to progress beyond the realm of instrument construction.

Under Instruments, the Sockloff paper attempts to delve into the logical considerations underlying the construction of a faculty evaluation instrument. This paper attempts to develop a skeleton for a model of learning and teaching for the purposes of constructing faculty evaluation instruments and to discuss some of the pitfalls that have been so carelessly neglected in the construction of instruments. Willard Warrington's paper gives us a good idea of the history of the approach in developing a popular instrument, while Donald Hoyt's paper takes a more practical, not-so-psychometric, and slightly humorous view of what is actually involved in attempting to intelligently construct an instrument. Richard Perry and Reemt Baumann show us that no matter how carefully an instrument is designed and constructed, the problems in its usage may be insurmountable.

Last, since studies of correlates of evaluation in faculty and student characteristics have tended to concentrate on confounding factors, the results of these studies are suggestive of the varieties of confounding factors vitiating the validities of the instruments that were used in these studies. While Wilbert McKeachie's paper summarizes a potpourri of results in this broad area, Jerry Gaff's paper suggests the shortcomings of traditional approaches to faculty evaluation. Gaff uses his results to support his views on education and related considerations in faculty evaluation. Mary Jo Clark and Robert Blackburn make use of a theoretical model from another field in their study of faculty characteristics.

Traditionally, prefaces are optimistic and sometimes laudatory about the contents of the prefaced volume. The break with tradition in this volume is meant more to encourage, rather than discourage, future work in this area. If the concept of having students evaluate their teachers is to be taken seriously, it is clear that some significant improvements are needed in this field. Apparently, faculty evaluation is a game that can be played by anyone, where demagoguery is easily mistaken for wisdom. If this condition is the result of the faddish nature of the area, and we know that fads live and die cyclically, hopefully something practicable will be learned before we experience its death and eventual revival. Although the papers contained within this volume represent some of the best work in this area, the reader should have little difficulty realizing that our present state of knowledge is rather rudimentary.

On a more positive note, many individuals have contributed to the success of this first conference, and thanks and appreciation is extended for their help. The Measurement and Research Center staff has given large amounts of time and good advice throughout all phases of the Conference. These individuals are:

Harold C. Reppert, Ph.D., Director

Abraham A. Panackal, Ph.D., Director of Achievement Testing

David D. S. Poor, Ph.D., Director of Statistical Data Analysis

J. Porter Tuck, Director of Educational Research

Edward Lake, Data Reports

Terry Sendrow, Information Systems

Estelle C. Kalstein, Office Manager

Posey Schwartz, Convention Secretary

In addition, Millard E. Gladfelter, Chancellor of Temple University, entertained us as banquet speaker, allowing us a respite from the long involved sessions; and Earl J. McGrath, former Director of the Higher Education Center at Temple University, willingly volunteered to moderate the Discussion of Issues and did so with aplomb under somewhat chaotic conditions.

Alan L. Sockloff

STUDENT EVALUATION OF FACULTY: WHY? WHAT? HOW?

Paul L. Dressel

Michigan State University

I should confess in the beginning that I am not greatly enthused about discussion of systematic evaluation of teaching by students in isolation from other approaches to evaluation of faculty services. I favor student evaluation, but I think that evaluation of faculty services is complicated and that too frequently ventures into the student evaluation of classroom teaching become simply a way of evading the broader problem of careful evaluation of all faculty activities. One of the common complaints about colleges and universities is that research is given prime consideration in the reward system and that little or no attention is given to teaching. Actually, I believe there are relatively few institutions in the country which systematically evaluate the research output of faculty members. I have known many faculty members who were promoted and given salary increases largely because of their published research, even though many of their associates (in private) expressed doubts of its worth or quality. There are only a few institutions that regularly collect and submit to scholars in other universities the research output of a person before a major promotion or the granting of tenure.

Faculty members commonly engage in student advising, and there is general complaint from both students and administrators that faculty advising is grossly inadequate. Nevertheless, too little has been done to collect systematically student appraisal of advising and even less has been done to improve faculty advising. Yet in most institutions any attempt to provide other systems of advising are thwarted by the insistence of the faculty that this is their prerogative, although usually their insistence is based on a false conviction that the student's major is the most important factor in his undergraduate program. I doubt that the student majoring in a discipline must be advised by faculty members in that discipline, and I know this arrangement does not insure good advising.

Faculty members also engage in extensive service on and off the campus. Some participate heavily, and even excessively, in committee work and in quasi-administrative work. These services are no more adequately evaluated than teaching. Why, then, should the pressure be on teaching rather than on the full range of faculty services when evaluation by students is discussed? First, students often complain about teaching; hence many persons-- including most of the faculty--feel that some opportunity should be provided for students to present their point of view. Second, the ready availability of the classroom and large numbers of students involved in classes make it relatively easy to use a few minutes of classroom time to collect a large number of reactions to the course and the teacher. Third, the development of objective formats--that is to say, a series of statements to which students can respond by checking some alternatives--makes it possible to collect a very large amount of data and process it speedily through use of electronic equipment. The net result of these three considerations is that student evaluation of teaching is undoubtedly the most prominent and the most discussed means of evaluation of faculty services. Numbers and the pseudo-objectivity of the responses give many people a sense of false security about the reliability and validity of the results; yet one has only to note that, in an objective format, students respond only to items included to realize the limitations of this approach. My observations on many campuses indicate that many of the more revealing statements which ought to be in such a form are excluded by the faculty as irrelevant to their conception of teaching responsibilities.

At best, most of these evaluation forms focus on what goes on in the classroom and on what the faculty member does in the way of clarifying objectives, making specific assignments, preparing examinations, giving grades, and the like. This is certainly not the whole of an individual faculty member's performance, and it does not even include all of his instructional contributions. There have been individuals whose classroom performance was abominable, but who have written excellent and widely used textbooks. Individuals adept at preparing tests and other evaluation materials may markedly affect the teaching of many members of the staff, yet not excel in the particular kinds of behavior usually involved in student evaluation forms. The faculty, too, may be very effective with some students while quite ineffective with others. As I recall my undergraduate days as a major in mathematics, I reconfirm my conviction of that time that most of my undergraduate teaching was bad, and that the mathematics teaching was deplorable. I did have two professors who were very effective in my particular case. Both of them, in effect, said that I was wasting time in the class and would profit more from independent work. One professor went so far as to guarantee an "A" in the course whether I did anything more or not. In both cases it was a welcome and beneficial release for me, and I really didn't lose much time sympathizing with those students required to attend class.

I conclude, then, from observation, experience, and some research, that evaluation of teaching by students is based on a very limited conception of faculty services and, especially or particularly, on a limited conception of the teaching act itself. The dangers inherent in this approach are that this involvement in evaluation may have more read into it than it deserves and that the involvement in time and resources may effectively eliminate any possibility of a broader evaluation. This last issue deserves more consideration, and I shall return to it later.

Some Questions About Student Evaluation

Several different questions about student evaluation need to be considered. First of all, what aspect of faculty performance do we want students to evaluate? Much of the answer depends upon how we interpret the pronoun "we." We, as faculty, usually wish the items in any evaluation form to be specific to the course, the content of that course, and our personal conception of the teaching act. Broader behavioral objectives definitive of a liberal education are generally rejected by the faculty. Usually the faculty do not want students to evaluate advising because they feel that advising is an extra duty thrust upon them for which there is no possible recognition or reward. In their advising, they are primarily concerned with majors and really have no interest in the broader aspects of advising that the undergraduate may find of great concern to him.

Likewise, faculty reject the idea that students can evaluate the quality and fairness of an examination or the justification of specific course requirements. Administrators, accustomed to hearing students complain about unreasonable assignments, poor examinations, inability to hear the professor, professorial absenteeism, and the like, generally take a broader point of view of what might be evaluated by students.

Students themselves generally take a rather narrow point of view. They are concerned that the professor express himself clearly, that his statements be audible, that his assignments be clear and not too demanding, that his examinations be directly related to classroom coverage, and that they neither require unreasonable memorization nor extensive thought. Students like some clarification of objectives, but are readily satisfied with a statement of the content to be covered and the requirements to be met in terms of examinations, papers, and the like. They are not encouraged to think about a course or the instruction as relevant to some of their personal interests or their other courses. They are not urged to view the course in terms of its contribution to a liberal or general education. Students don't really expect that, as a result of a particular course, they will be increasingly capable of independent effort in the type of materials studied in the course. In short, we impose such limits on what students evaluate that the student sees each course and each instructor in isolation rather than as a part of a much broader and more significant cumulative educational experience. Generally, students are being asked to evaluate petty details which have little significance to them and often no significance to the instructor who might wish to use the student reactions to improve his teaching. For example, I submit that when students in large numbers assert that "objectives are not clear" instructors obtain little assistance in how to improve the situation. When many students say that "not much was gained by taking this course,"I know that most instructors assume that this response is characteristic of students who get low grades, although it may as well characterize the views of those who get "A's." I find it singularly unhelpful to learn whether a group of students believes an instructor was friendly to students. The best teacher that I ever had was distinctly not friendly to students, although he wasn't unfriendly or antagonistic; he was simply a busy man and impatient with any delay or interference. He obviously spent many hours of time preparing for his classes, he carefully read any examinations or papers, and he was deeply concerned that his students learn something of significance. He did know more about his students than most of them suspected, but he was never characterized as friendly.

When students indicate that too much outside reading is required, one can scarcely judge whether this is a commendation or a criticism. Most of my own graduate students will respond in this manner to my two seminars when they compare those seminars with others that they have taken. On the other hand, they unanimously agree in a final assessment of the benefits gained from the seminars that the reading has been valuable. Students are frequently asked to respond to such an item as "the laboratory was a worthwhile experience." I have long since become convinced that most of the laboratory in freshman science courses is a waste of time and money, particularly when compared with alternative patterns of experience which might provide greater benefits. The freshman laboratory typically does not provide any vision of what scientific experimentation is all about; it's largely a cookbook and time-consuming procedure which fails miserably to educate the freshman student as to the nature of scientific exploration. Yet I agree with the faculty that most of the students are incapable of this judgment. Those who are would hesitate to record it in the face of the teacher's commitment to the laboratory.

Students are capable of evaluating much more than we permit them to do about evaluation of faculty effort. On the whole, they evaluate what we let them evaluate, and the faculty members tend to eliminate or ignore any aspects of student evaluation that might materially change the prevalent faculty conception of teaching.

What is good teaching? A simple answer is that good teaching produces effective learning, but that leaves open a wide range of views as to what constitutes good teaching. The individual who teaches mathematics as an end in itself follows the textbook and presents to the students a series of problem types. Generally speaking, he assumes that the students cannot read the material in the textbook which was rewritten by a professor to impress other professors rather than for the students. Hence, the teacher uses classroom time to make an exposition of the theory and work a number of problems of the same type. Ultimately, the examination samples these various problems and permits the student to earn his grade by demonstrating that he can indeed do what he has been asked to do from day to day. Seldom does he understand the theory which was developed for background in solving the problems. He may not have the least idea of their utility. The likelihood is that within a few months after completing the course he (unless he continues in mathematics) will have little recollection of the materials covered, less of what to do with particular problems, and almost no sense of the nature of mathematical reasoning and its widespread application in other fields.

Good teaching in the eyes of many faculty members is simply coverage of particular materials demanding certain knowledge and skills and testing to see that these have indeed been acquired for the moment. The development of broader abilities, attitudes, and insights which might enable the person to apply something of what he has learned to pursue independent study in the field--these and other broad liberal education outcomes are ignored. I am reminded of an individual who, by most standards, must be regarded as having been a very capable professor and dean who wanted help in evaluation of a freshman course, but rejected any attempt to state explicit objectives on the ground that the course was a first course which prepared to take a second and the second prepared to take a third, etc. until finally, if a person took enough courses in that particular discipline, he might be capable of doing something with it.

In short, my major concern about the typical approach to student evaluation of faculty is that it is ultimately dominated by a very inadequate conception of teaching and learning. At best, professors present a little better and students temporarily learn a little more of material which has limited, if any, long-term significance. The usual approach, which starts with students who know no better and works through faculty who studiously avoid any approach which would require a broader conception of the teacher-learning process, means that we simply reconfirm what exists. I sincerely doubt that teaching has been very much improved on any campus by the use of student evaluation forms. If their advent is marked by insistence that these become available to chairmen and deans, the battle lines are clearly drawn and ultimately the faculty will revoke that requirement. If the evaluation is optional with the faculty members, or at least optional in terms of their revealing it to chairmen, deans, or others, they will generally use it only to the extent to which complimentary reactions by their students are passed on for whatever benefits may be accrued while other reactions are ignored as irrelevant or as beyond the capability of student judgment.

The Process of Student Evaluation

The objective teacher rating form is so extensively used because of its convenience that other means of involving students in evaluation of teaching are overlooked. Any instructor, seriously concerned about his teaching, can learn much by careful observation of his students, by interviews with individuals, by classroom discussions, or by requesting essay comments to several questions at the end of examinations. Students may be reluctant to express some of their concerns directly to the instructor, but this in itself constitutes an evaluation of great significance. The instructor who cannot convince his students of his ability to separate his evaluation of student performance from student evaluation of the course or of his own performance has thereby identified a major deficiency. Until and unless he can tolerate frank discussion and criticism, he is unlikely to improve.

Yet students who are, on the whole, charitable in their appraisal of teaching may be unwilling to express their most critical concerns directly to an instructor. They may be even less willing to do so with departmental chairmen, instructors' colleagues, or deans. An expert interviewer, evaluator, and observer can bring out views and behavior not readily expressed or apparent to the teacher himself because of preoccupation with his own activities. My own experiences with such classroom observation convince me that few professors can appraise the quality of a discussion or are even aware that, in what passes for a discussion, they may talk for 40 or 45 minutes out of 50. I have, incidentally, verified this by use of a stop watch!

Some professors who reject objective check-lists and other objective formats are willing to use open-ended essay responses to questions or to a suggested list of course factors or characteristics. I rather like the critical incident approach or a request for comment on the best and worst aspect of a course. These do not lend themselves to generating norms. This is an advantage, in my judgment, for if evaluation is to be focused on improvement, evidence that an individual teacher is above or below average is not only irrelevant, but it may so affect the individual that he will not strive to improve. If already well above average--why bother? Seek rather for a raise or a promotion. If below average, an injured ego may indeed seek retribution on the students or undertake to discredit the entire system of evaluation. Teaching, like learning, is a very personal experience. Norms are no more conducive to improving teaching than to improving learning.

I have visited campuses in which students are encouraged to write letters, fill out forms, visit the dean, or in other ways present their complaints (or commendations) about teachers. The sampling here may be of concern to some, and the motivation of some of those using this approach may be suspect. But the extent to which such letters are written and the nature of the complaints registered involve some student behavior beyond the quiescent response to a form passed out in the classroom.

There are other aspects of student performance which are relevant to evaluating teaching. The extent to which students elect a course or a particular faculty member surely indicates evaluation of the worth of that experience. If common examinations of any kind are used, either for a course or some group of courses taught by the same individual, the examination performance of the students is certainly an evaluation of the teaching, although one must hasten to add that high performance on the examination has to be weighed in reference to the nature of the examination itself. Personally, I should not regard as an excellent teacher a professor whose students all made high grades on a very factual examination, although I know some faculty members who would be delighted by that evidence. Neither would I be happy with a high level of forced performance which resulted in avoidance of the field thereafter.

One aspect of student evaluation that interests me greatly and which is, I think, done the least is that of investigating changes in student behavior outside of the class and in following years. Some years ago I found on a college campus several groups of students in their senior year who were meeting bi-weekly to talk about developments in the natural sciences. These sessions had started spontaneously in the freshman year because of a course required of all students as one of the general education group. This course dealt in part with current developments in the sciences, and students became aware of certain kinds of magazines and reports, and they banded together for meetings to read and discuss these. Several of these were continuing three years later. I can think of nothing more potent in evaluaitng the effectiveness of a professor than the stimulation he provided for a group of students to continue their interests in an area originally forcibly brought to their attention by a freshman requirement.

But in a broader sense, if we have not, in teaching a course, given an individual some ideas, some techniques and insights of ability to do independent study on his own in that area, we have really given him nothing of significance. And it is generally the failure to deal with these broader behavioral outcomes that leaves me relatively cold to the usual practices in student evaluation.

Incidentally, some institutions have undertaken evaluation of teaching by alumni. I have some doubts about this approach because a few years after leaving college a student will have had such a variety of experiences that his recollection of contacts with specific instructors and courses as an undergraduate is likely to be far from accurate. Furthermore, there is a tendency in retrospect to see one's experiences through rose-colored glasses and perhaps to become more charitable of professorial weaknesses simply because of becoming aware of the extent to which people generally perform less effectively than might be desirable.

Uses and Benefits of Student Evaluation

In this section I propose to raise the general question of why we should encourage student evaluation of teaching. And again the answers are somewhat different depending upon our interpretation of "we." Students who become interested in some rating and reporting on faculty, at least in my experience, seem to be motivated largely by two considerations. (1) They have had some unfortunate experiences and, in some sense, they would like to record somewhere their dissatisfaction. (2) They hope also that by this means they might warn other students to avoid certain courses or instructors. Beyond this, some students hope that, by the publication of reports which reveal the poor quality of teaching, the reward system will be brought to bear upon these people, forcing them to improve or leave. I have no adequate basis for assessing the impact of student-conducted evaluation and reporting.

My impression, however, is that the magnitude of student interest in such surveys is far less great than the student initiators thought would be the case, and I am generally convinced that the impact of these published reports on the faculty is minimal. Even as a visitor to campuses using this approach, I have been distressed by some of the statements which have been published, particularly about young faculty members or teaching assistants for whom some alternative method of pointing out attention to weaknesses should have been used rather than a published report. And, although on the whole I have felt students were charitable in their interpretations, the sheer inexperience of students in evaluation and their lack of understanding and lack of sensitivity exhibited by some of the students in writing about the teaching of individual professors lead me to question the worth of such enterprises. Evaluation of teaching is a complex and difficult task.

A second possible use of student evaluation is with reference to the reward of faculty members and the assignments which are given them. Students would like to have something to say about promotions, granting tenure, and possibly the granting of salary increases or other forms of recognition to individuals. Many of them feel, with some reason, that reports on the quality of teaching ought to be used to eliminate or to reward professors rather than simply be collected in the vain hope that individuals will be inspired to improve their teaching. In many respects, I agree with the students, although I have seen more faculty members antagonized by student reports of inadequate teaching than I have who were motivated to improve. Indeed, I have seen few departments in which a significant proportion of the staff felt any confidence in their ability to appraise the teaching of the associates and, considering the lack of adequate means of appraisal, I tend to be quite skeptical of departmental assessments of good teaching. For example, I recently visited an institution with a Doctor of Arts program under way. Members of the department reported on how they were going to evaluate the required internship of each D.A. candidate. I was given some reports from recent visiting committees. These tended to criticize the novices for overly informal classroom procedures--sitting on the edge of the table, leaning against the wall, etc., as against apparently the faculty preferred stance, front and center with manuscript or notes on a podium. Another common critical comment about the intern's teaching procedure was the inadequacy of the lecture, its organization, or its depth. I also noted criticisms of certain aspects of lectures as indicating that the intern was not sufficiently sensitive to the underlying facts in some of his statements. Out of this came the recommendation that the student be required to take one or more additional graduate courses so that he could be more precise in his treatment of these matters. I doubt that teaching will be much improved by this approach. If departmental faculties really understood good teaching, we would have less of a problem with inadequate teaching. As it is, a new degree may not improve the situation. The Ph.D. surely does not train people for teaching and, if most of our faculties have no conception of teaching except that of the scholar delivering well-organized packages of knowledge to his students, improvement may be difficult via a new degree.

Quality of teaching should be a major factor in the reward system, but I do not believe that student ratings of teaching are an adequate basis for doing this, nor am I sanguine about many colleges or departments having a sufficient number of professors with a well-thought-out conception of what good undergraduate teaching is to feel sure that we can readily introduce any system capable of recognizing and rewarding good teaching. And furthermore, it is significant that, in collective bargaining, as it has developed in public schools and now gradually expands in higher education, the tendency is to avoid any approach which depends upon merit or good teaching. My own interpretation of this is that most faculty members are, first, not willing to admit they are not good teachers; second, they cannot admit, or will not admit, that there is sufficient agreement on what constitutes good teaching to pick out individuals for special recognition; and third, especially in colleges and universities, if students were carefully selected in the first place because of their enthusiasm for learning there would be no need for concern about good teaching.

Another reason for student evaluation projects is found in the research interests of some faculty members (often psychologists working with a sophomore sample). I have read much research on the qualities of good teachers and on the effectiveness of different methods of instruction. The cumulative impact of all of this research essentially is nil insofar as providing any guidance about how to improve teaching. The generalizations are suspect and of little use, for improving teaching is ultimately the process of working with individuals. I recall being told years ago in an education course that the use of sarcasm by a teacher was quite undesirable. I was immediately led to think about a number of professors whose gentle use of sarcasm needled students to think more deeply about an issue. This is only a simple example of someone's attempt to devise (by rationalization or research) a general and apparently reasonable principle of very limited validity. Prof. McKeachie argues that there are some general statements which can be made about the effectiveness of various methods of instruction. But with all deference to my good friend, I continue to doubt that we know anything about the relationship of any generalized method to specific outcomes. In the first place, I have grave doubts about studies which characterize relationships between methods and outcomes. In most cases, when I have looked at them closely, I have found that the so-called methods were not clearly defined or consistently used, and were often contaminated by other factors. Earlier I mentioned a case of a teacher who had talked for 45 out of 50 minutes. At the close of the class he remarked to me that this was the best discussion session he had had in some time. In another study, I found the majority of a control group regularly meeting with the experimental group because they found the latter's experiences were more exciting than their own.

A second problem is that the differences in method should be related to the objectives that the professor has in mind. I find few professors deeply concerned about objectives involving personal development, affect, values, or even the development of increasing independence and self-direction. In a study last year, I found one professor giving a lecture three times a week to a student enrolled in independent study. Yet both the professor and the records characterized the student's experience as independent study.

I do not object to research on the nature of teaching and learning. In fact, we need much more fundamental research than we have, but I would point out that research and evaluation are very different things. Research, in the long run, may provide us some insights from which we can move toward improvement; but the concerns of students and of critics of higher education are that we do something about improving teaching right now. This is evaluation.

Certainly from the point of view just mentioned, and probably from the point of view of this conference, improvement of instruction and of learning represent the two major concerns which justify evaluation by students. We need to note that in this process of improvement of instruction there are some problems which, in effect, negate improvement. Evidence will not improve instruction if that evidence is also used to deny a salary increase, negate a promotion, or decide a tenure action. Somehow, these administrative decisions and the process of improvement have to be separated from each other. For new instructors joining a staff, emphasis can well be on improvement of instruction with enough time lapse so that if two or three years later it becomes clear that the individual will not or cannot improve then appropriate action can be taken. If collection of data on the quality of teaching becomes available only at a point in time when a decision is to be made, then most faculty members will only resist, fight, and attempt to deny the validity of any undesirable information which accrues.

Any attempt to relate evaluation to the improvement of instruction and also to decisions about individuals will generate real difficulties. As has been true in so many cases, the attempt to develop an evaluation scheme involving student response generates a faculty demand that this be handled as a confidential feedback to individuals who may or may not see fit to share the results with others. This leads to a pattern of optional reporting or consultation in which individuals utilize only so much of an evaluation as they find suitable to their purpose. I have in a few cases learned of at least a temporary situation in which reports were placed with a department chairman and the faculty member was asked to sit down with the chairman for a formal discussion of the student ratings. I would have a great deal more confidence in this if I felt that most department chairmen were sensitive to what good teaching involves. Required consultation would be helpful if it could be used as the starting point of a program gauged to the needs of the individual professor and if it could help him, over time, improve the quality of his teaching and finally culminate in another reporting which would demonstrate that improvement. Those universities which have been able to set up a program of services to help professors analyze their course, course objectives and materials, to develop new materials, make use of educational technology, and other means to improve the quality of the learning of students have, I am convinced, done a great deal to improve the quality of teaching and learning. The difficulties I see here are twofold. First, there are not large numbers of professors who take advantage of these services and, if more were encouraged to do so, the costs might readily become prohibitive. Second, observation over a period of time indicates to me that individuals who become deeply concerned about their teaching and take advantage of all of these possibilities tend shortly to become involved in other activities. They may become administrators, they may become involved in committee work, sometimes they become consultants on these matters, and end up by retreating to a lower quality of instruction simply because they become so much engrossed with other matters or have moved to new assignments as a result of the venture into improvement of teaching and learning.

I would make another remark about encouraging faculty to look at their teaching. At the present time, when recommendations are made on faculty members, we usually lack the information required to determine whether a person is a good teacher or not. The individual, backed by his fellow faculty members, insists that if there is no evidence to demonstrate that he is an inadequate teacher then we must assume that he is a good teacher. And so we do. We could change the situation by informing everyone who joined the faculty as an instructor or assistant professor that he would not be promoted or given tenure until he provided convincing information about the outstanding quality of his teaching. In short, throw the burden back on the individual and then make available to him the help and the services to gain that information. I have not yet seen any institution that was willing to take this approach and, as collective bargaining becomes more widely prevalent, it may become impossible.

Another use or benefit of student rating is to alleviate student concerns and perhaps develop some good will by giving students an opportunity to participate in faculty appraisal. In some institutions students actually sit on committees with faculty in passing judgments on promotions, salary increases, and tenure. At this point, I am sure the student voice must have come impact. I doubt, however, that the usual student evaluation has any impact on the departmental recommendations with regard to individuals. Thus, in a sense, we gull the students into believing that their voice is heard, but actually ignore it, except that student appraisal of teaching does at least tend to promote faculty awareness of student reactions.

Possible Detrimental Effects of Student Evaluation

In accordance with my attempt to analyze the benefits of student evaluation, I should also consider the possible detrimental effects. One major point that I have already made is that the usual approach to student evaluation involves much too limited a conception of teaching. This limited conception of teaching has a two-way impact. On one hand, it allows the student to continue to think that teaching can be evaluated primarily on what goes on in a classroom situation. My own commitment is that teaching is more properly evaluated by the inspiration which it gives to the student to carry on his learning beyond the classroom situation. A second and related concern is that student evaluation, in the usual pattern, deals in generalities which have little to do with good teaching. The opportunity of a student to react to a statement that the objectives of the course are not clear may say something about the statement of objectives either in the departmentally-prepared syllabus or in the lecture as prepared by the instructor. In either case, it says very little about the appropriatness of these objectives or the extent to which the objectives carry beyond the specific content covered to the development of general abilities, insights, and values. The indication that the laboratory was or was not a worthwhile experience tells, at best, from a rather limited student point of view, whether the laboratory experience seemed worthwhile. The student has no basis for determining whether the laboratory was as effective as some other experience might have been, and he certainly has no basis for weighing the costs of the laboratory against possible demonstrations of some of the ideas conveyed through the laboratory experience. Such statements as this have very little directly to do with good teaching, and they provide no information which can be used as a basis for improvement. Most students may, given the statement that the instructor did or did not synthesize, integrate, or summarize, will respond to this in unsatisfactory or meaningless ways. If the instructor regularly, at the end of each class, attempts to summarize what he has covered, the students will probably recognize this. Nevertheless, that attempt to synthesize, integrate, or summarize may be grossly inadequate in terms of the immediate material and even less adequate in terms of the long-term development of concepts and principles in the course. In short, the fact that the instructor is noted as summarizing does not at all mean that he summarized well. Earlier we noted also that the ease with which student evaluation on a mechanical basis can be carried out makes this a very popular approach. At the same time, the involvement of time and energy in this approach becomes an excuse for not going any further in the evaluation of teaching and learning. In a sense, the detrimental effect here is that we yield to criticism by doing something, but choose to do something that is inadequate to correct the real problem which generates the concern. Putting it another way, we react as little as possible.

There is a need for balance in evaluation, and balance must be interpreted to include many things. The adequacy of the classroom situation itself needs to be evaluated. If too hot, too crowded, or too noisy, attention and learning will suffer. The objectives need to be examined in some depth. Many courses, especially in colleges and universities, have no formal statements of objectives, but simply assume that the materials covered are objectives in themselves. The objective is to cover the material without thinking through or really being concerned about the results in terms of new insights and abilities on the part of the individual student. The student is examined on how much of the material he has memorized. When objectives are unclear or inadequate, evaluation concentrates on the process. But improvement of the process is impossible unless based upon improvement in learning with regard to objectives. If these are regarded as inadequate by qualified observers, improvement is not possible. What the faculty member does (which is a part of the process of education that goes on in the classroom) and what the faculty member expects or requires of his students outside of the classroom also should be related to objectives. The culminating aspect of evaluation is always with regard to learning by the students. What have they achieved? And at this point, evaluation must not focus simply on what they have achieved in terms of the originally stated objectives, but also in terms of other by-products, side issues which may not have been contemplated. A student may have indeed met certain required knowledge goals in a course, but come out disliking the area so much that he vows never to have further contact with that field. In this case, the significance of the unanticipated outcome completely negates the actual gains made with regard to this specified outcome.

Cost Benefit Analysis

At this point, I shall undertake to draw together several strands of thought to deal with the general question, does student evaluation achieve benefits in proportion to the costs in time, energy, morale, dollars? We should note that many student evaluation programs require the use of a class period or part of a class period. What is intended to be 10 or 15 minutes for a response to a form often, by student contrivance, extends to 30 or 40 minutes. Even if the student is asked to take the form home to respond to it (at the risk of reducing the response total and polluting the response by discussion with roommates) some class time is usually required for passing out forms and explanation. But generally speaking, the amount of committee and administrative time involved in the preparation of a student rating form is the most expensive aspect of the whole process. My own experience indicates that faculty members are likely to insist that any evaluation form be thoroughly reviewed by a local committee, which probably means several tryouts, an extensive amount of work by some staff members, and a great deal of editorial work and elimination by the committee. The instrument coming out of a university committee usually is, by faculty insistence, circulated to departments for reactions, with the result that many of the more significant items (at least in my estimation) have been eliminated as irrelevant. No sooner is the instrument given than there are faculty criticisms and a demand for elimination of certain items and for review and revision of others. Typically, we have found at Michigan State that some faculty resist and will refuse to use any instrument in which statements are made to which students are expected to react. They demand an open-ended form with a series of questions to which students write an essay response. Their insistence that no objective format is meaningful in providing specifics for improvement is one with which I sympathize. I cannot avoid noting also that, in providing an essay response, the student almost totally negates any attempt to summarize student reactions in the form of norms.

In addition to these costs in time (which are seldom estimated), there are cash outlays for printing, scoring, and compiling norms. There are further staff time involvements in the many consultations with individuals within departments, with various committees, and the like. In any large university, I am quite sure that any careful assessment of the costs of student teacher rating forms would be of the magnitude of $5,000 or $10,000 per year. And in those years (probably every two or three) in which a major revision is required, the total costs, including all of the time of the many persons who become involved, may well run to $40,000 or $50,000. The question, then, that one has to weigh is whether the gains by the expenditure of funds in this way are justified in terms of the benefits gained. If I were to summarize the benefits of student ratings as I have seen them operate at Michigan State and other institutions where I have consulted, it would be as follows. First, the involvement of students in rating faculty is evidence of concern about the quality of teaching. Second, administrative support of such student ratings and financial support for the total process indicates an administrative position which favorably influences the student, although it may be rejected by the faculty as an invasion of academic freedom and of departmental and college privileges. Third, the extensive discussions that are involved on any campus when the matter of student rating of teaching is under consideration probably have some educational value. Members of the committees and others who become involved are led to think through some of the characteristics of good teaching, and this may have an influence on some of them transcending any direct benefits which come from the use of the forms which ultimately result. It would be very difficult to assess each of these educational benefits. My own observation leads me to believe that the discussions at the formative stage of such a program may be the most valuable result of the whole venture. Fourth, the development of a student rating project may affect hiring and reward criteria. I underline "may" because, in those situations where I have had any chance to observe, my conviction is that the lapse in time and the almost complete separation between programs of student rating and procedures for selecting new faculty make it very unlikely that there is anything more than the most general consciousness about teaching which carries over from the evaluation program to the selection of faculty. It has probably happened, but I have yet to learn of a faculty member who was asked to present student ratings on his teaching in applying for a position elsewhere.

My tentative conclusion from this review of student rating of teaching are the following:

1. The usual faculty and student conceptions of the nature, objectives, and obligations of teaching and learning (bound by traditions and limited by experience and bias) simply do not provide an adequate basis for student evaluation of teaching.

2. Unless based upon a conception of objectives and of teacher

obligations beyond the traditional classroom, the impact of student evaluation is very limited. It may indeed be more of a distraction than a benefit.

3. Student evaluation alone, whether by structured inventory or other means, is obviously not an adequate basis for judging total faculty effectiveness. It is also inadequate for assessing teaching effectiveness. Hence, unless balanced by other evidence, reliance on student evaluation may be both inequitable and dangerous.

4. Published student evaluations are not very useful to faculty members, are probably used by a relative minority of students, and they may be grossly unfair to junior members of a faculty whose careers are still in a formative stage and who should be receiving concrete positive help in improving their teaching rather than published criticisms made by naive individuals whose own conception of teaching, formed as it has been by their college experience, is grossly inadequate.

5. Finally, this paper has emphasized that there are other forms of student evaluation and rating scales, and that there are many other aspects of evaluation of faculty services which have some relationship to teaching. My own conviction, then, is that, in any institution in which there is concern about faculty performance, those involved in developing an evaluation program should think through in the broadest terms the obligations and activities of faculty and attempt to develop a complete evaluation system. After this has been done, several different ventures may be developed in terms of evaluation of aspects of faculty performance. I'm certain this will result in a realization that there are more facets and more interrelationships among these than student ratings can possibly provide. I believe that our approach to defining and collecting student ratings of teaching will be redefined if related to a broader concern about what faculty do and how well they do it.

THE STUDENT AS GODFATHER?

THE IMPACT OF STUDENT RATINGS ON ACADEMIA

John A. Centra

Educational Testing Service

Most of you, I'm sure, are familiar with the Godfather role made popular by the very successful book and movie. He was depicted as someone with a great deal of power over people and viewed by most with a mixture of awe, fear, and respect. In fact, his "offers that one could not refuse" were indeed, as some of you will recall, quite compelling.

There are some who fear that the college student, by virtue of the apparent increasing emphasis on student ratings of professors, could become the "Godfather" of the academic community. More exactly, they fear that too much emphasis could be put on these ratings and that, generally speaking, the power that students might acquire would not be in the best interest of the academic community.

These Cassandras can, in fact, point to the medieval universities as an example of unreasonable student influence over teachers. As Hastings Rashdall tells us in his writings about the medieval European universities, students at the University of Bologna not only paid teachers a "collecta" or fee (which apparently was determined by a teacher's ability to haggle), but they also could report teacher irregularities to the rector. For example, law texts were divided into segments, and each instructor was required to cover a particular segment by a specified date; to enforce this statute, the rector appointed a committee of students to report on dilatory professors, who were then required to pay a fine for each day that they had fallen behind.

While few people would take seriously the possibility that students are on the verge of assuming the role they played in medieval days, some do question the ultimate impact of student evaluations on teaching and learning. I will be more specific about some of their reservations later in this paper. In addition, I plan to discuss evidence of the positive effects of student ratings, and finally, since the impact of student ratings on certain aspects of academic life is not totally known, I will speculate about some possible consequences.

I've grouped my comments within five categories and will discuss the impact or possible impact of student ratings on the individual instructor, on teaching generally, on students, on administrators, and on the college.

The Individual Instructor

First, let me begin by discussing the person the ratings are meant to influence most: the individual teacher. There has been a good deal of skepticism over how much effect the ratings actually have on changing or improving instruction--particularly when the results are seen only by the individual teacher. Faculty conservatism, when it comes to educational changes, has been a well-known tendency, although there are signs that it may be less true now than in the past. For example, I recently had occasion to look at the responses of some 2800 college teachers to the question, "When did you last make changes in the teaching methods you are using?" About a fourth indicated that they had never made changes. On the other hand, about half said that they had changed their methods during the past two years. So it looks as if we should not indict all college teachers with the time-worn stereotypes of stodginess and traditionalism. Many apparently are willing to change their methods.

The question, though, is what causes teachers to change and, more germane to my topic, can ratings by students lead to any noticeable changes among college teachers? While a few investigators have noted that the ratings that teachers receive seem to improve over time, we know that we cannot assume a cause and effect relationship. Those changes could have been caused by any number of factors other than the initial student feedback.

One of the best ways to investigate the effects of student ratings on an instructor's practices is to employ an experimental design in which random groups of teachers receive feedback from students while other teachers--those in the control groups--do not. As some of you know I completed such a study within the past year with the cooperation of over 400 faculty members at five colleges. The details of that study are presented elsewhere (Centra, 1972), so I won't take the time to repeat them. But I would like to discuss briefly the results. The major conclusions of the study were, first, that changes in instruction (as assessed by repeated student ratings) occurred after only a half semester for instructors whose self-evaluations were considerably better than were their student ratings. If, in other words, teachers were especially "unrealistic" in how they viewed their teaching--unrealistic relative to their students' views, that is--then they tended to make some changes in their instructional practices, even though they had only a half semester to do so. I might add that such variables as the subject area of the course, sex of the instructor, and number of years the instructor had taught did not distinguish which instructors made changes; or to put it another way, none of the subgroups of teachers formed by these variables were more likely to change. The second conclusion was that a wider variety of instructors changed if given more than a half-semester of time and if they had some minimal information to help them interpret their scores. Let's consider briefly the implications of each of these findings.

Starting with the first result, why do you suppose changes in teaching procedures were related to the discrepancy between self-evaluations and student ratings? Actually this result was predicted at the outset of the study because there was fairly good reason to expect it, based on social psychological theory. As a matter of fact there are several similar theories that help explain the finding. Most are referred to as self-consistency or equilibrium theories, the central notion being that an individual's actions are strongly influenced by his desire to maintain a consistent cognitive condition with respect to his evaluations of himself. What this means is that when student ratings are much poorer than an instructor's self-ratings, a condition of imbalance (Heider, 1958), dissonance (Festinger, 1957), or incongruency (Newcomb, 1961; Secord & Backman, 1965) is created in the instructor. In an attempt to become more consistent, or in more theoretical terms to restore a condition of equilibrium, the instructor changes in the direction indicated by his students' ratings.

These theories assume, of course, that most instructors place enough value on collective student opinion, and that instructors know how to go about making changes. Undoubtedly some teachers merely write off student judgment as unreliable or unworthy, and for these individuals, changes are unlikely even though they may be called for. At least the changes are unlikely if the only motivation comes from within the individual teacher. Increasingly, however, student ratings of professors are becoming public information, and in these instances there is undoubtedly a good deal of social pressure to change. In fact, not only is there social pressure, but in some instances there is economic pressure, since the ratings may be used in salary and tenure deliberations. But as I've said, it is not always clear to the teacher how to change, if indeed he or she believes the change would be an improvement. And this leads me to the implications of the second finding from my five-college study.

I mentioned that with additional time and with some interpretative information, the ratings for a more diverse group of teachers had changed in a positive direction. Not surprisingly, many teachers need more time to change their procedures, particularly in those areas that cannot be quickly altered (clarifying course objectives, for example). Yet if student ratings are to have maximum impact, I believe we need to do more in interpreting the results to instructors and in helping them improve. One of the reasons that we need to help instructors interpret their ratings is that the ratings are typically skewed in a positive direction. Most of us already know this, but the average teacher does not. On a five-point scale, he views his mean score of 3.6 as above average, when actually it may well be only average or even below average if compared to other teachers. Parenthetically, I might add that instructor self-ratings, not surprisingly, are skewed even more positively than student ratings. And faculty peer ratings based on classroom visits, according to some data I've recently collected, are also generally more favorable than student ratings. In any event, some kind of normative or comparative data is important for interpreting student ratings, and, perhaps, the more the better. The instructor might be given the choice of comparing his students' responses to those of other teachers at his institution, or to those of members of his department; or perhaps he may prefer a more cosmopolitan comparison--such as to instructors from a sample of other institutions, or perhaps to a national sample of teachers in his field. The point is that a variety of comparisons might be made available to the instructor so that he can decide which are most meaningful.

Some of these comparison data are already being made available to instructors, though not always with the variety I've suggested. But I'm afraid that they do not totally solve the problem. There will still be some instructors who need special help, and for this reason Kenneth Eble (1971), for one, has suggested that individual instructional counseling be made freely available. A teacher counselor might not only help instructors interpret their student evaluations but could, of course, also suggest particular ways in which to improve. A few institutions are already doing this, but in these times of tight money this will probably remain a limited endeavor.

I'd like therefore to mention another possibility that I'm now pursuing. In place of an individual counselor I would propose substituting the next best thing: the computer. One of the remarkable feats of the computer is that it can be programmed to produce a verbal interpretation of a numerical summary. Rather than means, standard deviations, or percentile ranks, each professor could instead get several paragraphs of prose telling him how he differs from his own expectations and how he differs from some predesignated group, such as other teachers in his field. The number-leery professor need not worry about whether his scores are significantly different--the computer will make that interpretation. Moreover it would even be possible to refer the instructor to specific materials, books, or even video tapes pertinent to his weaknesses. For example, if students said his course objectives were not made clear, or if they rated the quality of exams poorly, there would be several excellent references dealing with these topics suggested to the instructor. In fact, there's really no need to rely on the computer to produce these suggestions--we ought to be doing that sort of thing right now.

Before moving on to discussing other categories, I'd like to make one last point regarding the effects of student ratings on the individual teacher. With the emphasis generally put on mean scores or percentile ranks of scores, I'm afraid that the individual teacher is being influenced to see his class only as a homogeneous glob. Anyone who has taught knows that quite frequently there are several types of students in the typical class, each of which may be reacting a little differently to the teacher and the course. These different types and their various viewpoints do not mean that the ratings are unreliable in the sense that there is a great deal of fluctuation or inconsistency in student responses. We know that student ratings are reliable, as indicated by the numerous intraclass reliability studies that have been reported. What I'm talking about is identifying subgroups of students who differ systematically in their ratings. Is there, in short, some rhyme or reason to the diversity of viewpoints that may exist in the typical class?

One way to investigate this question is to use factor analytic techniques that allow one to group individuals rather than items as is usually the case (see Tucker & Messick, 1963). The only study I have found that looked at this question had investigated students' general notions about types of teachers rather than their specific ratings of individual teachers (Rees, 1969). So I've undertaken some additional analyses--first with three large classes separately and then across a larger sample of courses--which indicate that there are frequently three or sometimes four points of view represented in a single class. Each of these groups sees various aspects of the course or the instruction they are receiving somewhat differently from the other groups. One group, for example, may have rated the instructor as generally ineffective, but at the same time indicated that the instructor was well organized and usually accessible; another group might have rated the instructor as ineffective and inaccessible. Unfortunately, I don't at this point have enough information about student characteristics that would allow me to describe the groups. Ultimately, however, it may be possible to alert the individual teacher to relevant subgroups or points of view in the class; these points of view might be identified by student characteristics information, or they might be identified by patterns of ratings. Until then, teachers should be encouraged to look at the distribution of student responses to the items on their rating form--and not only at the mean scores. While no one expects them to please all of their students all of the time, instructors ought to be aware of how they interact with different segments of the class.

Impact on Teaching Generally

Closely related to the effects of student ratings on the individual teacher is the possible impact that they have on teaching generally. The critics of student ratings claim that an undue emphasis on the ratings, such as using them to assist in decisions on faculty promotions, can have adverse effects on instruction. What are some of these adverse effects? First, some critics claim that the ratings do not allow for individual styles of teaching, that they instead force everyone to be measured on the same yardstick. Few people would try to assess artists or composers on the same yardstick, according to one skeptic of student ratings. That skeptic goes on to say, in an article in The American Scholar, that:

> The art critic need not evaluate portraits painted by Picasso, Whistler, and Rembrandt in terms of criteria for effectiveness common to all three. He finds it possible to examine each artist's work in terms of the artists' own goals, or to identify the strengths and weaknesses of an individual painting in terms of relations of parts to the whole [Kossoff, 1972, p. 89]

Even though I don't happen to believe that teaching and art are entirely comparable, we know enough about teaching to know that individuals can have quite different styles, and that they should probably develop the style that best fits their personality and approach. I'll return to this point in a minute.

A second adverse effect of student ratings, according to the same critics, is that they encourage traditional modes of teaching. Most rating forms are indeed directed at classes taught in some combination of lecture-discussion, but logically so--that happens to be the way most courses have been taught and the forms are merely reflecting what is typically the case. The question is, however, are other methods such as student-centered learning, or nondirective teaching, or team teaching being stifled by the typical student rating forms? The answer, in my opinion, is that they are if an institution does not allow some flexibility in the application of student ratings. This means that for some courses, and this is still a relatively small number on most campuses I suspect, it is necessary either to supplement or disregard items in the traditional rating forms.

Flexibility in the employment of student ratings is, in other words, extremely critical. Many of the widely used forms have been developed through what might be called the consensus approach. In other words the developers have asked samples of faculty members (or faculty members and students) to identify specific characteristics that are important in teaching. Those areas or items for which there was the greatest consensus were then included in the rating instrument. Generally speaking, the items have centered around such factors as course organization, teacher-student interaction, and communication or verbal fluency. It's clear that this approach does not produce an instrument that reflects any particular theory of teaching. And that probably has made good sense in view of the fact that it would be difficult to get any college faculty to agree on a single theory of teaching.

While most forms allow individual instructors to add their own items to a basic set, there are other ways in which the rating forms can be even more flexible. If the items are to be used in making decisions on faculty members, then the individual teacher might be allowed to eliminate those items that are not relevant to his style. Better yet, a system might be implemented which allows teachers to both choose and weigh in advance the items which they feel most adequately reflect their style of teaching and what they are trying to accomplish in the course. At least one institution is now working on such an approach.

Impact on Administrators

Another group that student ratings influence--albeit more indirectly than previous groups--are college administrators. I have two observations to offer regarding this. First, that in instances where the ratings are used in making decisions on promotions, it may well be that the dean or department chairman's job becomes a little easier.

National surveys have told us that frequently the judgments of one or more administrators are relied on to assess teaching effectiveness, particularly at smaller colleges. Not many people would defend this as a very wise or valid approach. If we can assume that the evidence provided by student evaluations means not only wiser decisions but also ones that are more easily defended, then students' evaluations make the administrators' jobs easier and more effective. Some, I realize, would debate that point.

A second observation that I have is that student evaluations may well be contributing to what seems to be a current groundswell for administrator evaluations by faculty members. A not too infrequent request to ETS is for an instrument to evaluate administrator performance. Apparently the feeling is that if faculty can be evaluated by their constituents, then by all means so can administrators. Increasingly, it would appear that they are. For example, the trustees of the State University of New York announced in January that the presidents of the 29 colleges operated by the state will have to undergo intensive evaluation of their records every five years. But I'm not at all sure that a handy-dandy machine-scored instrument could be developed that would measure reliably and validly an administrator's performance. More likely the charge is for administrator accountability (to use the still-currently "in" word), in which an individual is accountable not only to his superiors but also to his subordinates.

Impact on Students

According to the results of the ACE 1972 annual survey of freshmen, students feel generally that faculty promotions ought to be based in part on student ratings. That opinion was endorsed by three-quarters of the students from the 373 institutions in the survey. This probably comes as no surprise. The past decade has, of course, been a time when students have demanded a greater role in institutional decision-making, and the evaluation of teaching would appear to be an area in which they feel they can make a unique contribution. Where student ratings have been incorporated into faculty evaluation procedures, therefore, the impact on students is likely to be quite positive; at least each of them can feel that he or she is helping the institution make important educational decisions. This is not to be taken lightly. While in the past teachers and administrators have been willing to give students a say in such areas as the establisment of student personnel policies and regulations, they've been more reluctant to relinquish their hold on academic decision-making.

Aside from this, probably the major impact of student ratings on students is provided by published course and teacher critiques. While some institutions make public the results of college-sponsored student evaluations (and some publish course guides based on detailed descriptions provided by the instructor), most of the critiques are based on surveys that are student initiated and conducted. As you might suspect, these student-produced critiques vary considerably in quality from one institution to another; in fact, they may vary from year to year at single institutions, depending on which students get involved. The worst of the critiques have been based on poor samples and frequently border on sensationalism by highlighting the juiciest of criticisms. Needless to say these critiques do neither the teachers nor the students who purchase them much good. But what about the better publications; what about the critiques based on thorough methodology and which, as in some instances, also give the teacher an opportunity to respond to his student evaluations? Do they have a suitable reason for being? One might argue that they provide information that the college catalog or other publications don't provide and this would seem to be a valid purpose. Nevertheless there are many faculty members who object strongly to student-conducted course ratings. Their objections have been delineated by Kerlinger in a 1971 article in School and Society. He argues that student initiated ratings result in "instructor hostility, resentment, and distrust," and thus alienate faculty members from their work. He goes on to suggest that ratings are legitimate only if conducted voluntarily by professors and used for self-improvement. Obviously then, not only is there concern for who initiates and conducts a student rating of instruction program, but also to what end the results are to be used.

Needed, it seems to me, is a major study of the effects of student ratings when they are used to assist in deciding whom to promote. There are a number of questions that such a study might investigate. For example, to what extent do faculty become alienated? Which types become most alienated? Does it encourage traditional teaching and limit teaching styles, as already discussed? Does it erroneously reinforce the notion in students that the instructor is largely responsible for how much students learn in a course? This last point may be true regardless of how student rating results are used and in spite of the fact that many of the rating forms ask students about their own effort and involvement in the course. But the major question to be answered by such a study is whether more defensible promotion decisions are made when student evaluations are included as part of faculty assessment.

Impact on the College

The last category that I will comment on is the impact, or possible impact, of student ratings on the college.

I've already discussed changes that take place among individual teachers--or at least among some teachers. But can an institution, or perhaps the departments within an institution, learn something about themselves from student evaluations? A corollary question is: "What can the institution or department then do about what they've learned?"

Let's start at the department level. A seldom mentioned, though seemingly worthwhile, use of student ratings is that of providing departments with information about the effectiveness of their offerings as seen by students. To do this it would be necessary to combine the ratings of all members in a department, and items dealing with specific as well as general course objectives should be included in the assessment. In addition to these course-instructor evaluations, a sort of major field questionnaire might be given to seniors. Princeton University, for one, has been using a major field or department questionnaire for the past several years. While not the typical application of student evaluations, the assessment of departmental offerings would seem to be worthy of consideration by other institutions.

Another point that might be made concerning the departments is that, as many of us have discovered, there are some interesting variations in the evaluations that teachers in different subject fields receive. Among a group of some 450 teachers, for example, I found that courses in the natural sciences, relative to those in humanities, social sciences, and education and applied subjects, were seen by students as having a faster pace, as being more difficult, and as being less likely to stimulate student interest. In addition, teachers perceived the natural science teachers in the sample as less open to other viewpoints. Humanities teachers, in comparison to those in the other three general subject areas, were less likely to inform students of how they were to be evaluated, and there was less agreement between the announced objectives of humanities courses and what was actually taught.

The obvious question is whether it is the subject matter itself that produces these differences or the types of individuals within each of the subject areas. It may well be a combination of both. At any rate, patterns of ratings would indicate that subject fields or departments might focus on certain apparent weaknesses (for example, humanities professors might attend workshops on improving their evaluation procedures).

The whole notion of focusing on weaknesses highlighted by student evaluations could be applied at the college level even more generally. If a college is able to compare itself to other colleges--that is, if the aggregate ratings of all teachers can be compared--then it may be possible to identify specific weaknesses. Workshops in that particular aspect of instruction might then be offered to assist in faculty improvement.

Conclusion

In this paper I've attempted to discuss the effects or possible effects of student evaluations on academia. It has been apparent throughout the discussion that the major effects are to a large extent, dependent upon how the ratings are used. Their primary uses can perhaps be summarized best by adapting Michael Scriven's (1967) terms for the two major functions of tests: formative and summative evaluation. Tests used formatively, according to Scriven, give the instructor periodic feedback on his students' progress, thus telling the instructor what needs to be stressed in the future. The summative function of tests, as the term implies, is a way of providing a summative evaluation of each student at some point in time.

When student ratings of instruction are used formatively--that is, when they are used by instructors as a source of feedback on their teaching--the evidence indicates that some changes are made by the instructor. And most likely we can improve on this with better interpretation of the results. The effects of using student ratings in a summative way--that is, in making administrative decisions on faculty--is a little more difficult to assess. As a researcher I feel we ought to learn more about the side-effects. But if I were a department chairman or dean faced with increasingly tougher tenure-promotion decisions, or if I were a faculty member who felt that his teaching was not being rewarded, then I might hold a different view. Certainly student evaluations are no less trustworthy than other methods now available to assess teaching performance, and when combined with other methods, they probably contribute to a fair judgment.

In closing, I'd like to return briefly to the title of this talk. As you have realized by this time, I don't believe that students, through student ratings, are or will become the Mario Puzo type of Godfather to the academic community. But this is not to say that they might not function is a limited way as proper Godfathers. Traditionally, of course, a Godfather has had a much more positive image; he essentially is one who helps provide guidance and direction to those in his charge. While I'm not suggesting that students are the new saviors of academia, or that college teachers must rely on the guidance of their students, I do think that a well-designed student ratings program can do more to benefit than to harm the academic community.

References

Centra, J. A. The utility of student ratings for instructional improvement. Project Report 72-16. Princeton, N. J., Educational Testing Service, 1972.

Eble, K. The recognition and evaluation of teaching. Project to improve college teaching, Salt Lake City, Utah, 1971.

Festinger, L. A theory of cognitive dissonance. Evanston, Ill.: Row, Peterson, 1957.

Heider, F. The psychology of interpersonal relationships. New York: Wiley, 1958.

Kerlinger, E. Student evaluation of university professors. School and Society, October 1971, 353-356.

Kossoff, E. Evaluating college professors by "scientific" methods. The American Scholar, Winter 1972, 79-93.

Newcomb, T. M. The acquaintance process. New York: Holt, Rinehart and Winston, 1961.

Rees, R. D. Dimensions of students' points of view in rating college teaching. Journal of Educational Psychology, 1969, 60 (6), 476-482.

Scriven, M. The methodology of evaluation. American Educational Research Association monograph series on curriculum evaluation, No. 1, Perspectives of curriculum evaluation, 1967.

Secord, P. F., & Backman, C. W. An interpersonal approach to personality. In B. A. Maher (Ed.), Progress in experimental personality research, Vo. 2. New York: Academic Press, 1965.

Tucker, L. R., & Messick, S. An individual difference model for multidimensional scaling. Psychometrika, 1963, 28 (4), 33-367.

THE USEFULNESS OF STUDENT EVALUATIONS IN IMPROVING COLLEGE TEACHING

Lawrence M. Aleamoni
University of Illinois

In the past few years as a result of the 1970 student strikes and the emphasis on accountability, course and instructor evaluation has been placed in the spotlight. In an attempt to build a total instructional evaluation system, a great deal of emphasis has been placed on student evaluations of course and instructor. In order for student evaluations to be considered an integral part of a total instructional evaluation system, they must be both reliable and valid.

Of the various systems developed for student evaluation of course and instructor, the Illinois Course Evaluation Questionnaire (CEQ) has perhaps the most extensive reliability and validity data to support it as well as the most extensive norm data base. Norm data have been collected continuously since 1966 at the University of Illinois, Urbana-Champaign campus. The CEQ is used to collect student attitudes towards a course and instructor and its purpose is to enable faculty members to collect evaluative information about their teaching. Once the instructor has used the CEQ and submitted the forms for analysis, two copies of the results are returned only to the instructor. As the number of measures on each course is increased, it becomes possible to obtain a relatively stable indication of the difference between courses. This aids in the interpretation of the actual differences between an obtained section score for a particular instructor and the average scores for all the sections represented in that course.

The analysis of item inter-relationships and the subscore inter-relationships indicated that no one element, related to a course, disproportionately influenced the students' evaluation of the course (Spencer & Aleamoni, 1969). It appears that there is a "general course attitude" cultivated by the student as he is exposed to previous student's comments, the instructor, the testbook, the course, etc., and this is the framework from which he responds when answering the CEQ items.

It would seem, on the basis of three validity studies (Stallings & Spencer, 1967; Swanson & Sisson, 1971; Aleamoni & Yimer, 1972), the face validity of the CEQ, and its high reliability, that extremely low scores on a particular subscore should indicate problem areas in an instructor's teaching procedure. Whereas, stable high scores should point to an effective instructional program as viewed by students. All available validating evidence (both published and unpublished studies), to date, indicates that the CEQ does indeed identify courses that are considered to be excellent or poor.

After using the CEQ, the instructor receives results (see Appendix A) which allow him to compare his course item means to institutional course item means (via deciles) and his course subscale means to norm subscale means categorized by (a) rank of instructor, (b) level of course, (c) institution, (d) college, and (e) all institutions that have used the CEQ throughout the United States. The subscale results allow the instructor to obtain an indication of major areas of strengths and weaknesses in the course. Once the areas of weakness have been identified by the subscales, then looking at the item results helps to focus on the more specific problem areas. The CEQ items are completely diagnostic but do serve to elicit diagnostic responses from the instructor teaching the course. It provides a means whereby some evaluation of the teaching process can occur; other means can be arranged and are available such as asking more diagnostic questions in the optional item section available on the CEQ form, or having peers sit in on actual class sessions, etc. It is important to recognize, however, that student opinions are in existence and do affect learning--and they do provide a source of quite reliable and valid data relative to the effectiveness of instruction (Costin, Greenough & Menges, 1971).

In order to provide instructors with items that may be more relevant or diagnostic for their particular courses, a catalog of items was generated by the Measurement and Research Division of the Offices of Instructional Resources at the University of Illinois, Urbana-Champaign campus. The items were gathered from all existing sources such as institutional, national, departmental, and individual instructor questionnaires. They were then restated so that the response categories of strongly agree (SA), agree (A), disagree (D), and strongly disagree (SD) would apply. This then made it possible for those items to be used in the "Optional Item" section of the CEQ (see Appendix B).

This collection of some 270 items was divided into 19 categories consisting of: (a) instructor contribution, (b) attitude toward students, (c) student outcomes, (d) relevance of course, (e) use of class time, (f) organization and presentation, (g) clarity of presentation, (h) instructor characteristics, (i) interest of presentation, (j) expectations and objectives, (k) behavioral indications of course attitude, (l) general attitude toward instructor, (m) speed and depth of coverage, (n) out-of-class, (o) examinations, (p) visual aids, (q) grading,

(r) assignments, and (s) laboratory and recitation.

The response to the availability of the catalog of optional items was gratifying in that it was not finished until December 12, 1972, less than four weeks before the end of the fall semester. Of 1414 course sections using the CEQ during fall semester 1972, approximately 313 made use of the optional item section.

After the instructor has decided to use the CEQ and/or any optional items of his choice, it is then up to him to decide what to do with the data. If he feels that the interpretation manual (Aleamoni, 1972) and abbreviated interpretation sheets are not sufficient to help him identify areas that may need improvement in the course, he can then arrange for a conference with one of the members of the Measurement and Research Divison staff. Such a conference would begin with a close scrutiny of the CEQ subscale results to see if any problem existed based on the norm data available. If a problem area was identified (such as Method of Instruction) then a close look at the items making up that subscale would be in order. If, in the discussion with the instructor the source of difficulty is identified, then the discussion would shift to possible ways of trying to resolve the difficulty. If, on the other hand, the source of difficulty cannot be identified using the existing items and the instructor's recall, then procedures (such as the use of optional items that are much more diagnostic) would be explored to be able to identify the specific problem.

It has been through a process such as this that instructors have been able to use student evaluations to identify instructional problems and then rectify them. Obviously, the success or failure of such a venture rests solely with the instructor and his willingness to both gather and use the data provided him.

A question that naturally arises from the above considerations is, "Can student evaluations of instruction and instructor be useful in improving college teaching once they are made available to the instructor?" Although there has been a great deal of anecdotal evidence to suggest that such evaluations do have a positive effect, no studies to date were available to support that "evidence." Since the author has been involved in utilizing student evaluations to help instructors identify and diagnose instructional problems, the data was available to conduct the present study.

Method

Instructors at two different institutions (University of Arizona at Tucson and Sheridan College at Sheridan, Wyoming) who had used the CEQ during the fall, 1971 and spring, 1972 terms for their courses were the subjects of the present study. Each of these instructors was then scheduled to talk with the author about his/her results. The conferences were conducted individually at the home campus of the instructor and took approximately 15 to 20 minutes. The conference began with a close scrutiny of the CEQ subscale results to see if any problems existed based on the norm data available. If a problem area was identified (such as Method of Instruction) then a close look at the items making up that subscale would be in order. If, in the discussion with the instructor the source of difficulty was identified, then the discussion shifted to possible ways of trying to resolve the difficulty. If, on the other hand, the source of difficulty was not identified using the existing items and the instructor's recall, then procedures (such as the use of optional items that are much more diagnostic) were explored to be able to identify the specific problem.

In order to attempt to answer the question of usefulness of student evaluations in improving college teaching, each instructor who had participated in the individual conferences was subsequently followed-up to see if any significant change had occurred in their student ratings in subsequent terms in the same or continuous courses. Similar CEQ data for instructors who were not able to participate in the individual conferences was available to use as a control group measure.

Means, standard deviations, class sizes, and norm deciles were obtained for each of the above instructors on five of the CEQ subscales as well as the Total. That data (presented in Table 1) was then analyzed to determine if the conferences had any significant effect in helping the instructor improve his/her teaching as reflected in subsequent student evaluations measured by the subscales and Total score of the CEQ.

Insert Table 1 about here

Results

In looking at the norm decile changes that took place for the lowest subscale value discussed in the conference (see Table 2), it appears that the conferences did have a significant effect especially when compared to the control group norm decile changes. The average norm decile increase for the experimental group as observed in Table 2 is 3.94 compared to .57 for the control group. It varies slightly for each of the two institutions. The range of norm decile increase for the experimental group is from 2 to 8 compared to from -2 to 3 for the control group.

Insert Table 2 about here

References

Aleamoni, L. M. Results and Interpretation Manual: Illinois Course Evaluation Questionnaire, Research Report No. 331. Urbana, Illinois: Measurement and Research Division, Office of Instructional Resources, University of Illinois at Urbana-Champaign, 1972.

Aleamoni, L. M. & Yimer, M. An investigation of the relationship between colleague rating, student rating, research productivity, and academic rank in rating instructional effectiveness. Research Report No. 338. Urbana, Illinois: Measurement and Research Division, Office of Instructional Resources, University of Illinois at Urbana-Champaign, 1972.

Costin, F., Greenough, W. T. & Menges, R. J. Student ratings of college teaching: Reliability, validity, and usefulness. *Review of Educational Research*, 1971, *41*(5), 511-535.

Spencer, R. E. & Aleamoni, L. M. The Illinois Course Evaluation Questionnaire: A description of its development and a report of some of its results. Research Report No. 292. Urbana, Illinois: Measurement and Research Division, Office of Instructional Resources, University of Illinois at Urbana-Champaign, 1969.

Stallings, W. M. & Spencer, R. E. Ratings of instructors in Accountancy 101 from video-tape clips. Research Report No. 265. Urbana, Illinois: Measurement and Research Division, Office of Instructional Resources, University of Illinois at Urbana-Champaign, 1967.

Swanson, R. A. & Sisson, D. J. The development, evaluation, and utilization of a departmental faculty appraisal system. *Journal of Industrial Teacher Education*, 1971, *9*(1), 64-79.

Table 1

Means, Standard Deviations, Sample Sizes, and

Norm Deciles for CEQ Subscales

BEST COPY AVAILABLE

Institution Instructor	Experimental Control	Pre Post	N	Data Type	CEQ Subscales					
					General Course Attitude	Method of Instruction	Course Content	Interest Attention	Instructor	Total
Arizona 1	Experimental	Pre	20	Mean	2.95	2.59	2.92	2.64	2.94	2.81
				S.D	.79	.81	.73	.85	.72	.79
				Norm Decile	3	3	4	3	2	3
		Post	18	Mean	3.49	3.25	3.22	3.11	3.25	3.25
				S.D.	.39	.37	.36	.61	.55	.38
				Norm Decile	8	8	8	8	0	8
Arizona 2	Experimental	Pre	22	Mean	3.45	3.15	3.19	3.30	3.20	3.22
				S.D.	.52	.63	.57	.52	.59	.59
				Norm Decile	8	9	7	9	5	7
		Post	20	Mean	3.72	3.44	3.36	3.53	3.43	3.44
				S.D.	.55	.65	.68	.59	.68	.68
				Norm Decile	9	9	9	9	8	9
Arizona 3	Experimental	Pre	13	Mean	3.19	3.01	2.90	2.81	3.29	3.03
				S.D.	.48	.51	.59	.71	.63	.61
				Norm Decile	5	7	4	4	6	5
		Post	14	Mean	3.13	2.95	3.00	2.95	3.22	3.03
				S.D.	.47	.64	.56	.68	.51	.56
				Norm Decile	5	6	6	6	5	6

Table 1 (cont.)

BEST COPY AVAILABLE

Institution Instructor	Experimental Control	Pre Post	N	Data Type	CEQ Subscales					
					General Course Attitude	Method of Instruction	Course Content	Interest Attention	Instructor	Total
Arizona 4	Experimental	Pre	49	Mean	3.33	3.05	2.96	3.00	3.32	3.10
				S.D.	.60	.68	.66	.64	.66	.65
				Norm Decile	6	8	5	6	6	6
		Post	69	Mean	3.34	3.01	3.02	3.06	3.34	3.12
				S.D.	.60	.71	.64	.70	.65	.67
				Norm Decile	7	6	7	7	7	7
Arizona 1	Control	Pre	17	Mean	3.29	2.98	3.02	2.99	3.23	3.08
				S.D.	.54	.68	.64	.68	.64	.64
				Norm Decile	6	7	5	6	5	6
		Post	18	Mean	3.26	3.04	2.99	3.03	3.15	3.07
				S.D.	.50	.58	.57	.56	.55	.57
				Norm Decile	6	7	6	7	4	6
Arizona 2	Control	Pre	7	Mean	3.70	3.50	3.38	3.41	3.66	3.48
				S.D.	.46	.50	.56	.60	.51	.57
				Norm Decile	9	9	9	9	9	9
		Post	5	Mean	3.60	3.38	3.42	3.57	3.70	3.46
				S.D.	.44	.85	.43	.48	.33	.45
				Norm Decile	9	9	9	9	9	9
Arizona 3	Control	Pre	12	Mean	3.76	3.54	3.55	3.57	3.72	3.58
				S.D.	.45	.52	.58	.61	.52	.59
				Norm Decile	9	9	9	9	9	9
		Post	13	Mean	3.50	3.08	3.32	3.30	3.48	3.30
				S.D.	.55	.74	.36	.61	.33	.42
				Norm Decile	9	7	9	8	8	8

Table 1 (cont.)

Institution Instructor	Experimental Control	Pre Post	N	Data Type	CEQ Subscales: General Course Attitude	CEQ Subscales: Method of Instruction	CEQ Subscales: Course Content	CEQ Subscales: Interest Attention	CEQ Subscales: Instructor	CEQ Subscales: Total
Arizona 4	Control	Pre	12	Mean	2.91	2.22	2.73	2.31	2.97	2.63
				S.D.	.62	.57	.72	.70	.66	.71
				Norm Decile	2	0	2	1	2	1
		Post	8	Mean	3.14	2.59	2.98	2.80	3.16	2.90
				S.D.	.71	.86	.54	.60	.67	.62
				Norm Decile	5	2	6	4	5	4
Arizona 5	Control	Pre	23	Mean	3.32	2.98	2.74	2.98	3.19	3.00
				S.D.	.70	.65	.80	.77	.71	.75
				Norm Decile	7	6	3	6	5	5
		Post	29	Mean	3.31	2.94	2.86	3.02	3.01	2.98
				S.D.	.57	.60	.41	.57	.39	.42
				Norm Decile	7	6	4	7	3	5
Sheridan 1	Experimental	Post	9	Mean	3.18	3.33	2.89	3.19	3.34	3.15
				S.D.	.39	.41	.18	.38	.30	.26
				Norm Decile	5	9	5	8	7	7
		Pre	5	Mean	3.27	2.67	2.75	2.77	2.75	2.81
				S.D.	.60	.76	.67	.70	.71	.71
				Norm Decile	7	3	3	4	0	3
Sheridan 2	Experimental	Post	8	Mean	3.27	3.00	3.29	3.23	3.25	3.19
				S.D.	.37	.75	.45	.38	.36	.45
				Norm Decile	6	6	9	8	6	8
		Pre	16	Mean	3.06	3.20	3.14	3.05	3.20	3.10
				S.D.	.83	.80	.87	.82	.76	.82
				Norm Decile	4	8	8	7	5	7

Table 1 (cont.)

BEST COPY AVAILABLE

Institution Instructor	Experimental Control	Pre Post	N	Data Type	CEQ Subscales: General Course Attitude	Method of Instruction	Course Content	Interest Attention	Instructor	Total
Sheridan 3	Experimental	Post	11	Mean	3.38	3.32	3.35	3.01	3.45	3.31
				S.D.	.67	.64	.45	.69	.35	.47
				Norm Decile	7	9	9	6	8	9
		Pre	38	Mean	2.84	2.60	2.73	2.57	2.94	2.75
				S.D.	.73	.72	.63	.73	.71	.70
				Norm Decile	2	3	2	2	2	2
Sheridan 4	Experimental	Post	17	Mean	3.45	3.45	3.17	3.35	3.54	3.38
				S.D.	.48	.44	.50	.61	.36	.42
				Norm Decile	9	9	8	9	9	9
		Pre	22	Mean	3.10	2.59	2.70	2.78	2.89	2.83
				S.D.	.64	.68	.80	.71	.69	.73
				Norm Decile	5	2	2	4	1	3
Sheridan 5	Experimental	Post	13	Mean	3.38	3.19	2.91	2.83	3.32	3.12
				S.D.	.74	.77	.52	.65	.72	.63
				Norm Decile	7	8	5	5	7	7
		Pre	20	Mean	3.15	2.92	2.87	2.77	3.17	3.00
				S.D.	.66	.67	.74	.77	.77	.72
				Norm Decile	5	6	4	4	5	5
Sheridan 6	Experimental	Post	32	Mean	3.51	3.35	3.03	3.18	3.58	3.31
				S.D.	.48	.57	.39	.62	.31	.39
				Norm Decile	9	9	7	8	9	9
		Pre	14	Mean	3.24	3.25	2.91	2.90	3.36	3.13
				S.D.	.71	.53	.72	.83	.64	.70
				Norm Decile	6	8	5	5	7	7

Table 1 (cont.)

BEST COPY AVAILABLE

Institution Instructor	Experimental Control	Pre Post	N	Data Type	CEQ Subscales					
					General Course Attitude	Method of Instruction	Course Content	Interest Attention	Instructor	Total
Sheridan 7	Experimental	Post	10	Mean	2.97	2.63	2.70	2.47	3.10	2.78
				S.D.	.44	.44	.20	.60	.17	.25
				Norm Decile	3	3	2	2	4	2
		Pre	31	Mean	3.05	2.46	2.76	2.61	2.77	2.75
				S.D.	.73	.85	.76	.83	.82	.81
				Norm Decile	4	1	3	3	0	2
Sheridan 8	Experimental	Post	11	Mean	3.27	2.92	2.85	2.85	3.15	3.01
				S.D.	.38	.64	.28	.56	.50	.40
				Norm Decile	6	6	4	5	4	6
		Pre	12	Mean	3.30	2.77	2.71	2.69	2.96	2.90
				S.D.	.65	.73	.90	.90	.79	.81
				Norm Decile	7	4	2	3	2	4
Sheridan 9	Experimental	Post	8	Mean	3.67	3.73	3.22	3.58	3.73	3.54
				S.D.	.34	.33	.41	.33	.36	.30
				Norm Decile	9	9	9	9	9	9
		Pre	10	Mean	3.10	2.81	2.86	2.90	3.15	2.95
				S.D.	.44	.58	.47	.49	.53	.50
				Norm Decile	5	5	4	5	4	5
Sheridan 10	Experimental	Post	19	Mean	3.34	2.52	3.03	2.61	3.15	2.94
				S.D.	.38	.70	.27	.65	.59	.42
				Norm Decile	7	2	7	3	4	5
		Pre	28	Mean	3.31	2.70	2.94	2.62	3.17	2.90
				S.D.	.69	.75	.70	.83	.68	.77
				Norm Decile	7	3	6	3	5	4

Table 1 (cont.)

Institution Instructor	Experimental Control	Pre Post	N	Data Type	CEQ Subscales					
					General Course Attitude	Method of Instruction	Course Content	Interest Attention	Instructor	Total
Sheridan 11	Experimental	Post	9	Mean	3.54	3.00	3.43	3.40	3.44	3.33
				S.D.	.61	.96	.35	.72	.34	.51
				Norm Decile	9	6	9	9	8	9
		Pre	8	Mean	3.25	2.80	3.11	3.09	3.05	3.02
				S.D.	.50	.72	.51	.66	.68	.65
				Norm Decile	6	4	8	7	3	6
Sheridan 1	Control	Post	52	Mean	3.28	2.88	2.87	2.89	3.23	3.02
				S.D.	.54	.79	.42	.64	.49	.48
				Norm Decile	7	5	4	5	6	6
		Pre	13	Mean	3.45	3.13	2.81	3.29	3.47	3.20
				S.D.	.61	.75	.89	.68	.68	.76
				Norm Decile	9	7	3	8	8	8
Sheridan 2	Control	Post	8	Mean	3.70	3.31	3.34	3.63	3.61	3.43
				S.D.	.33	.82	.54	.42	.53	.50
				Norm Decile	9	9	9	9	9	9
		Pre	8	Mean	3.56	2.91	3.27	3.23	3.27	3.24
				S.D.	.50	.95	.70	.68	.60	.71
				Norm Decile	9	6	9	8	6	8

Table 2

Norm Decile Changes for the Lowest Subscales Value Discussed in the Individual Conferences

Institution	Experimental			Control		
Arizona	Pre	Post	Increase Decrease	Pre	Post	Increase Decrease
1	2	6	4	5	4	-1
2	5	8	3	9	9	0
3	4	6	2	9	7	-2
4	5	7	2	0	2	2
5				3	4	1
Mean	4.00	6.75	2.75	5.2	5.2	.00
Sheridan						
1	2	6	4	3	4	1
2	0	7	7	6	9	3
3	4	6	2			
4	2	8	6			
5	1	9	8			
6	5	7	2			
7	5	7	2			
8	0	4	4			
9	2	4	2			
10	4	9	5			
11	1	6	5			
12	3	8	5			
Mean	2.42	6.75	4.33	4.5	6.5	2.0

Appendix A

```
                 RESULTS FOR THE OBJECTIVE ITEMS ON THE ADVISOR QUESTIONNAIRE

20140 [redacted] EDPSY 490  SECTION H    ENROL=0005  FALL 1971      03620J

                                     SEX
                             FEMALE  MALE  OMIT
                               0.20  0.20  0.60

                                 MAJOR-MINOR
                          MAJOR MINOR OTHER  OMIT
                           0.40  0.20  0.40  0.00

                                COURSE OPTION
                                REQ ELECT  OMIT
                               0.40  0.40  0.20

                                  PASS-FAIL
                                YES    NO  OMIT
                               0.00  0.60  0.40

                                   STATUS
                      FRESH  SOPH    JR    SR  GRAD OTHER  OMIT
                       0.00  0.00  0.00  0.00  1.00  0.00  0.00

                               EXPECTED GRADE
                            A     B     C     D     E  OMIT
                         0.60  0.40  0.00  0.00  0.00  0.00

                                COURSE GRADE
                            A     B     C     D     E  OMIT
                         0.80  0.20  0.00  0.00  0.00  0.00

                              INSTRUCTOR GRADE
                            A     B     C     D     E  OMIT
                         1.00  0.00  0.00  0.00  0.00  0.00

ITEM    SA     A     D    SD  OMIT  BEST  MEAN  S.D.  DECL  0123456789
  1.  0.00  0.00  0.40  0.60  0.00    SD  3.60  0.55   9             *
  2.  0.00  0.00  0.40  0.60  0.00    SD  3.60  0.55   7           *
  3.  0.60  0.40  0.00  0.00  0.00    SA  3.60  0.55   9             *
  4.  0.40  0.60  0.00  0.00  0.00    SA  3.40  0.55   9             *
  5.  0.80  0.20  0.00  0.00  0.00    SA  3.80  0.45   8            *
  6.  0.60  0.40  0.00  0.00  0.00    SA  3.60  0.55   9             *
  7.  0.80  0.20  0.00  0.00  0.00    SA  3.80  0.45   9             *
  8.  0.00  0.20  0.40  0.40  0.00    SD  3.20  0.84   8            *
  9.  0.20  0.80  0.00  0.00  0.00    SA  3.20  0.45   8            *
 10.  0.00  0.00  0.40  0.60  0.00    SD  3.60  0.55   9             *
 11.  0.00  0.00  0.20  0.80  0.00    SD  3.80  0.45   9             *
 12.  0.20  0.80  0.00  0.00  0.00    SA  3.20  0.45   7           *
 13.  0.60  0.40  0.00  0.00  0.00    SA  3.60  0.55   9             *
 14.  0.00  0.00  0.40  0.60  0.00    SD  3.60  0.55   9             *
 15.  0.00  0.00  0.20  0.80  0.00    SD  3.80  0.45   9             *
 16.  0.80  0.20  0.00  0.00  0.00    SA  3.80  0.45   9             *
```

BEST COPY AVAILABLE

Appendix A (continued)

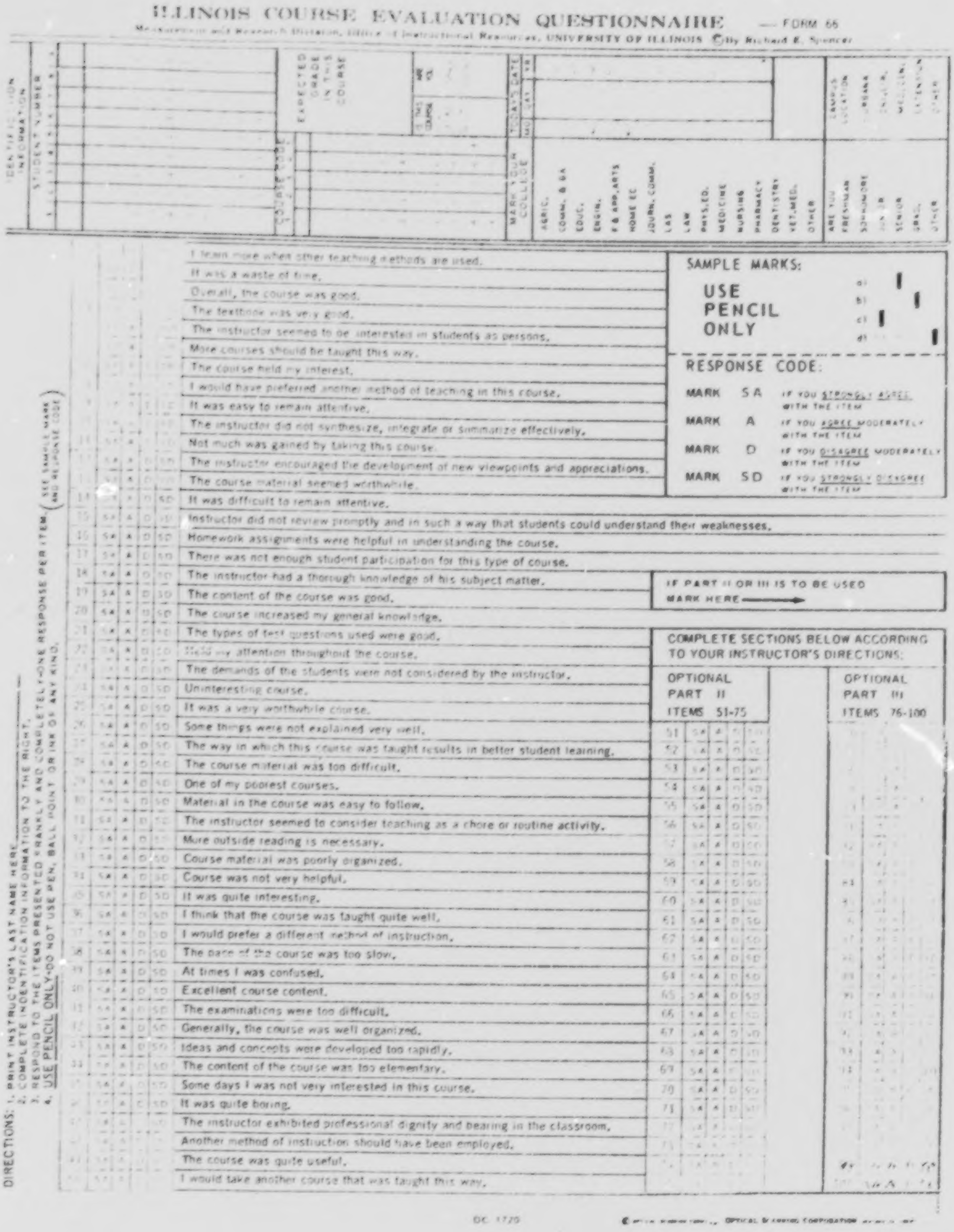
```
* * *   MERMAC -- TEST ANALYSIS AND QUESTIONNAIRE PACKAGE   * * *

ITEM    SA     A     D    SD  OMIT   BEST  MEAN  S.D.  DECL  0123456789
 17.  0.00  0.00  0.20  0.80  0.00     SD  3.80  0.45    9            *
 18.  1.00  0.00  0.00  0.00  0.00     SA  4.00  0.00    9            *
 19.  0.20  0.80  0.00  0.00  0.00     SA  3.20  0.45    9         *
 20.  0.60  0.40  0.00  0.00  0.00     SA  3.60  0.55    9            *
 21.  0.40  0.40  0.20  0.00  0.00     SA  3.20  0.84    8           *
 22.  0.40  0.60  0.00  0.00  0.00     SA  3.40  0.55    9            *
 23.  0.20  0.00  0.20  0.60  0.00     SD  3.20  1.30    6         *
 24.  0.00  0.00  0.20  0.80  0.00     SD  3.80  0.45    9            *
 25.  0.40  0.60  0.00  0.00  0.00     SA  3.40  0.55    8           *
 26.  0.00  0.00  1.00  0.00  0.00     SD  3.00  0.00    8           *
 27.  0.60  0.40  0.00  0.00  0.00     SA  3.60  0.55    9            *
 28.  0.00  0.20  0.80  0.00  0.00     SD  2.80  0.45    1    *
 29.  0.00  0.00  0.00  1.00  0.00     SD  4.00  0.00    9            *
 30.  0.00  0.80  0.20  0.00  0.00     SA  2.80  0.45    5        *
 31.  0.00  0.00  0.00  1.00  0.00     SD  4.00  0.00    9            *
 32.  0.00  0.40  0.60  0.00  0.00     SD  2.60  0.55    1    *
 33.  0.00  0.00  0.40  0.60  0.00     SD  3.60  0.55    9            *
 34.  0.00  0.00  0.20  0.80  0.00     SD  3.80  0.45    9            *
 35.  0.40  0.60  0.00  0.00  0.00     SA  3.40  0.55    8           *
 36.  0.60  0.40  0.00  0.00  0.00     SA  3.60  0.55    9            *
 37.  0.00  0.00  0.40  0.60  0.00     SD  3.60  0.55    9            *
 38.  0.00  0.00  0.60  0.40  0.00     SD  3.40  0.55    8           *
 39.  0.00  0.40  0.60  0.00  0.00     SD  2.60  0.55    6         *
 40.  0.40  0.60  0.00  0.00  0.00     SA  3.40  0.55    9            *
 41.  0.00  0.20  0.60  0.20  0.00     SD  3.00  0.71    6         *
 42.  1.00  0.00  0.00  0.00  0.00     SA  4.00  0.00    9            *
 43.  0.00  0.00  0.80  0.20  0.00     SD  3.20  0.45    8           *
 44.  0.00  0.20  0.40  0.40  0.00     SD  3.20  0.84    6         *
 45.  0.00  0.00  0.60  0.40  0.00     SD  3.40  0.55    9            *
 46.  0.00  0.00  0.20  0.80  0.00     SD  3.80  0.45    9            *
 47.  0.80  0.20  0.00  0.00  0.00     SA  3.80  0.45    9            *
 48.  0.00  0.00  0.40  0.60  0.00     SD  3.60  0.55    9            *
 49.  0.40  0.60  0.00  0.00  0.00     SA  3.40  0.55    8           *
 50.  0.60  0.40  0.00  0.00  0.00     SA  3.60  0.55    9            *

--SUBSCORE--         ITEMS  RESP  MEAN  S.D.   REL  RANK LEVEL INSTI  COLL  OVER-
                                                                            ALL
GENERAL ATTITUDE         8  1.00  3.65  0.48  0.90  NONE     8     9  NONE     5
METHOD                   8  1.00  3.55  0.55  0.85  NONE     9     9  NONE     9
CONTENT                  8  1.00  3.07  0.57  0.66  NONE     5     7  NONE     7
INTEREST                 8  1.00  3.55  0.50  0.93  NONE     9     9  NONE     9
INSTRUCTOR               8  1.00  3.67  0.62  0.00  NONE     9     9  NONE     9
SPECIFIC ITEMS          10  1.00  3.40  0.64  0.09  NONE     9     9  NONE     9
TOTAL                   50  1.00  3.48  0.60  0.93  NONE     9     9  NONE     9
```

BEST COPY AVAILABLE

Appendix B



ILLINOIS COURSE EVALUATION QUESTIONNAIRE — FORM 66

Measurement and Research Division, Office of Instructional Resources, UNIVERSITY OF ILLINOIS ©By Richard E. Spencer

IDENTIFICATION INFORMATION

STUDENT NUMBER

EXPECTED GRADE IN THIS COURSE

TODAY'S DATE MO. DAY YR.

CAMPUS LOCATION: URBANA, CHICAGO, MEDICAL, EXTENSION, OTHER

COURSE CODE

MARK YOUR COLLEGE: AGRIC., COMM. & BA, EDUC., ENGIN., F & APP. ARTS, HOME EC, JOURN. COMM., LAS, LAW, PHYS. ED., MEDICINE, NURSING, PHARMACY, DENTISTRY, VET. MED., OTHER

ARE YOU: FRESHMAN, SOPHOMORE, JUNIOR, SENIOR, GRAD., OTHER

DIRECTIONS:
1. PRINT INSTRUCTOR'S LAST NAME HERE
2. COMPLETE INDENTIFICATION INFORMATION TO THE RIGHT.
3. RESPOND TO THE ITEMS PRESENTED FRANKLY AND COMPLETELY-ONE RESPONSE PER ITEM. (SEE SAMPLE MARK AND RESPONSE CODE)
4. USE PENCIL ONLY-DO NOT USE PEN, BALL POINT OR INK OF ANY KIND.

I learn more when other teaching methods are used.
It was a waste of time.
Overall, the course was good.
The textbook was very good.
The instructor seemed to be interested in students as persons.
More courses should be taught this way.
The course held my interest.
I would have preferred another method of teaching in this course.
It was easy to remain attentive.
The instructor did not synthesize, integrate or summarize effectively.
11 SA A D SD Not much was gained by taking this course.
12 SA A D SD The instructor encouraged the development of new viewpoints and appreciations.
13 SA A D SD The course material seemed worthwhile.
14 SA A D SD It was difficult to remain attentive.
15 SA A D SD Instructor did not review promptly and in such a way that students could understand their weaknesses.
16 SA A D SD Homework assignments were helpful in understanding the course.
17 SA A D SD There was not enough student participation for this type of course.
18 SA A D SD The instructor had a thorough knowledge of his subject matter.
19 SA A D SD The content of the course was good.
20 SA A D SD The course increased my general knowledge.
21 SA A D SD The types of test questions used were good.
22 SA A D SD Held my attention throughout the course.
23 SA A D SD The demands of the students were not considered by the instructor.
24 SA A D SD Uninteresting course.
25 SA A D SD It was a very worthwhile course.
26 SA A D SD Some things were not explained very well.
27 SA A D SD The way in which this course was taught results in better student learning.
28 SA A D SD The course material was too difficult.
29 SA A D SD One of my poorest courses.
30 SA A D SD Material in the course was easy to follow.
31 SA A D SD The instructor seemed to consider teaching as a chore or routine activity.
32 SA A D SD More outside reading is necessary.
33 SA A D SD Course material was poorly organized.
34 SA A D SD Course was not very helpful.
35 SA A D SD It was quite interesting.
36 SA A D SD I think that the course was taught quite well.
37 SA A D SD I would prefer a different method of instruction.
38 SA A D SD The pace of the course was too slow.
39 SA A D SD At times I was confused.
40 SA A D SD Excellent course content.
41 SA A D SD The examinations were too difficult.
42 SA A D SD Generally, the course was well organized.
43 SA A D SD Ideas and concepts were developed too rapidly.
44 SA A D SD The content of the course was too elementary.
45 SA A D SD Some days I was not very interested in this course.
46 SA A D SD It was quite boring.
47 SA A D SD The instructor exhibited professional dignity and bearing in the classroom.
48 SA A D SD Another method of instruction should have been employed.
49 SA A D SD The course was quite useful.
50 SA A D SD I would take another course that was taught this way.

SAMPLE MARKS:

USE PENCIL ONLY

a) b) c) d)

RESPONSE CODE:

MARK SA IF YOU STRONGLY AGREE WITH THE ITEM
MARK A IF YOU AGREE MODERATELY WITH THE ITEM
MARK D IF YOU DISAGREE MODERATELY WITH THE ITEM
MARK SD IF YOU STRONGLY DISAGREE WITH THE ITEM

IF PART II OR III IS TO BE USED MARK HERE →

COMPLETE SECTIONS BELOW ACCORDING TO YOUR INSTRUCTOR'S DIRECTIONS:

OPTIONAL PART II ITEMS 51-75

51 SA A D SD
52 SA A D SD
53 SA A D SD
54 SA A D SD
55 SA A D SD
56 SA A D SD
57 SA A D SD
58 SA A D SD
59 SA A D SD
60 SA A D SD
61 SA A D SD
62 SA A D SD
63 SA A D SD
64 SA A D SD
65 SA A D SD
66 SA A D SD
67 SA A D SD
68 SA A D SD
69 SA A D SD
70 SA A D SD
71 SA A D SD

OPTIONAL PART III ITEMS 76-100

DC 1720

© [illegible] OPTICAL SCANNING CORPORATION [illegible]

FACULTY EVALUATION: SOME CONSIDERATIONS AND A MODEL

Kenneth O. Doyle, Jr.
University of Minnesota

Some months ago I happened to be having dinner with a fellow from the governor's staff "in one of our great midwestern states." The topic of universities came up in our conversation, particularly the topics of accountability and faculty evaluation. I was describing some of the problems involved in developing systems of faculty evaluation when he cut me off: There's nothing to it, he snapped; you simply assign monies to departments on the basis of their contribution to the gross national product!

I'm not going to tell you what happened after that-- just that it was not one of the most enjoyable meals I've experienced! His comment scared the daylights out of me, though, and underscored the importance of developing our own internal systems of evaluation before something less meaningful--and less palatable--is imposed on us.

With this added motivation I went into the literature with hopes of finding systems of evaluation that our institution might try on for size. I talked with faculty and students and administrators from various schools. What I found--with a few encouraging exceptions--was that faculty evaluation is a chaotic enterprise, as technically, politically, and conceptually complex as even the most masochistic of us could hope to enjoy.

Since I'm a bit of a compulsive sort, I needed to try to make order out of this chaos. Let me share with you what I've done thus far.

Considerations Concerning Faculty Evaluation

I believe there are a number of considerations that obtain for any system of faculty evaluation. We need to think about the purpose of the evaluation, the focus and consequences of the evaluation, sources of measurable data and the quality of those data. We need to attend to the goals of the institution. And we need to consider the media for gathering and reporting data and the temporal dimension along which the data must be gathered and interpreted. I'd like to develop each of these considerations a bit, then tie them together into the beginnings of conceptual schema, and finally show some applications of that schema. Although everything I say should pertain to all of faculty evaluation-- advising, research, governance, and service as well as teaching-- I'll draw most of my examples from the evaluation of teaching.

Purposes of Evaluation

There seem to be three more or less distinct and commonly proposed reasons for undertaking an evaluation: (1) to help improve faculty performance, (2) to help make personnel decisions concerning faculty; and (3) to provide a criterion measure for various kinds of educational research. Another purpose exists exclusively for the evaluation of teaching, namely to provide information that could help students choose their courses. Since I think that any criterion measure we might want to provide for research can come from purposes (1) or (2), I'll limit my remarks to the other purposes: to improve performance, to help in personnel decisions, and, for teaching only, to counsel students. Lets look at each in more detail.

Evaluation to improve faculty performance, which seems to be the most frequently stated purpose for doing evaluation, is distinguished from the other kinds of evaluation in that it attempts to diagnose strong and weak points in faculty behavior with the intent of helping remedy the weaknesses and reinforce the strengths. I want to emphasize that when I say "faculty performance" I'm not talking exclusively about teaching; I'm talking about the evaluation of all aspects of professional behavior--advising research, governance, and public service, as well as teaching. Nor am I restricting

BEST COPY AVAILABLE

our information to student data; I'm including data from colleagues, administrators, the public, and the faculty member himself.

Evaluations for personnel decisions focus on the rank, pay, and tenure determinations that lie near the heart of the student ratings controversy. These are evaluation data that help in the selection of faculty from a pool of applicants, in the placement of existing staff according to their abilities and attitudes (not just their interest and availability), and in the retention and promotion (or demotion) of faculty as a consequence of their professional performance. Again we need to remember that the sources of data are many and the behaviors to be evaluated varied.

People sometimes seem to make too clear-cut a distinction between these two purposes of evaluation. In theory, such a differentiation is sound, and it leads to some pointed considerations about, for instance, levels of reliability and validity that need to be established for the different uses of the data, and about techniques for gathering and analyzing information. (E.g., typical forced-choice scales are more suitable for personnel decisions than for improving performance because these scales don't usually furnish diagnostic or formative information.) But in practice the distinction breaks down to some extent. For example, although we might claim that the reliability of a particular instrument permits its use "only" for improving teaching, we have no way to restrict the use of the data once they are out of our hands. (Eventually I would hope that these two purposes will become even less distinct, that data to improve teaching and data for personnel decisions will overlap considerably more than they do now.)

The third purpose for evaluation seems to pertain only to teaching evaluation for the purpose of counseling students. These evaluations are intended to provide information that students might use to select among available classes or instructors-- or, for that matter, institutions.

This kind of information is by nature public and might be made available directly to students in bookstores, in student unions, in departmental or college offices, and so forth, or access to it can be restricted to certain professionals-- advisors and counselors, for instance. Other modes of access have been suggested. Some students at our institution have suggested a university telephone number at which a student operator could read the information to callers--rather a large responsibility for the operator and rather a busy operator at some times of the year! And a group of unusually imaginative students has been considering a system of computer terminals (CRT's) strategically located around the campus, which students could use to call up course-selection information from central data storage pools. More typical examples of this kind of evaluation are the phoenix-like Salvage from the University of Minnesota, the Advisor from the University of Illinois, and an intriguing two-part description/ evaluation handbook from the University of Utah that seems to avoid many of the problems inherent in these kinds of undertakings.

I think data for this purpose need some special scrutiny. There are the usual problems concerning the reliability and validity of published information, but the special problem here seems to be the General Bullmoose Fallacy that what's good for the average student is good for all students. I would be much more comfortable if published data were (almost?) exclusively objective descriptions of course goals, contents, and other characteristics, or—better still—if what the course offered were spelled out in terms of a profile of educational needs. Although this idea is not rare with regard to institutional profiles, little or no work of this kind seems to be taking place on the more specific classroom level.

But evaluation of teaching for purposes of course selection is probably here to stay. And so we have three kinds of evaluation that seem to cover what most of us mean by the term.

Dimensions of Faculty Activity

Once we know why we want an evaluation, we need to know what we want to evaluate. What aspect of the faculty member's behavior do we want information about? The major thrust of this conference has to do with student evaluation of teaching, but there are certainly other faculty activities that might profit from evaluation of some formal and systematic kind: advising, research and fundraising and publication, governance (e.g., committee work), public service, and so forth.

Each of these rather broad areas can be subdivided. For example, with regard to classroom teaching, focus might be on the objectives of instruction, the behaviors of the teacher or tutor (communication, organization, etc.), the various instructional materials (texts, other readings, handouts, audio-visual materials), the physical environment, and the social environment. To this listing we can add really anything that "impinges on the senses of the people involved", subject only to the constraints of manageable length and "reasonable" content.

Clearly I'm working toward a stimulus-organism-response conceptualization of the teaching process, and the list I've just described details to some extent the stimulus component. There is also the organism component, by which I mean the cognitive operations that the student applies to this stimulation. To evaluate a teacher by looking at the cognitive processes of students - cognition, memory, convergent and divergent thinking, and evaluating, to use Guilford's list - is theoretically possible, is probably of critical importance, but is certainly beyond our present capabilities. Nevertheless, this is a focus about which we need to be occasionally reminded. J.P. Guilford has furnished some of the classical work on cognitive operations, and Bloom has provided his taxonomy; but some of the most exciting and most recent work is being carried on by Russell Burris and associates at the University of Minnesota's Center for Programmed Instruction. In the context of computer-assisted instruction of material from beginning German to hematology to literary criticism and insurance law, Burris is working toward the identification, definition, and measurement of the dimensions of breakthroughs here. For the time being, however, I'm afraid that the inner workings of the student are beyond our reach.

But there is still the other side of the stimulus-organism-response structure, the response or output or product or performance side, which is essential to an evaluation of teaching. What did the student get out of the course? What student products or performance can we look at as indices of the effectiveness of the teacher? In the usual classroom situation, we can look at term papers, quizzes, and examinations. We can listen to oral reports and give oral exams. We can observe demonstrations. And we can evaluate work samples, whether the work is a statue in a studio arts class or criticism of a research design in a measurement class. The point is that we need to analyze products or performances from the student if we want to claim even a relatively comprehensive system for evaluation of the teaching component of faculty behavior. The fact that propels me so forcefully to this emphasis is not the aliberal vocational training argument but the human need to be goal-oriented. I worry that most of our evaluation activities pertain to the input side of the S-O-R structure - our own teaching behaviors, the materials we use, the social and physical environments in which we teach. I contend that more emphasis on the student response side would help disengage us from too much preoccupation with ourselves, our "styles", and our materials and would lead us to focus on those goals toward which our efforts are intended. Furthermore, this goal-orientedness should make any stylistic changes we make more likely to be valid in the sense of contributing to student learning.

In this vein, I would like to suggest a somewhat more orderly than customary approach to examining students, both for the sake of the testing itself and for the sake of that part of faculty evaluation that depends on student performance. I'd like to enter a plea for planned examinations, tests explicitly constructed according to a schema that reflects the purposes of the instruction and that recognizes not only the differential importance of the various subtopics of the material but that tests students on different "epistemological" levels - recall of fact, comprehension of ideas, application, analysis, synthesis. Bloom's Handbook and Thorndike's instant-classic on Educational Measurement would be superb reference works in this regard.

But back to the evaluation of faculty. Obviously we can't judge a teacher on the basis of unqualified student performance. We need to attend to complex qualifiers like student ability and motivation and other factors that I'll mention under the heading of Quality of Data.

Quality of Data

The quality of all evaluative information is critically important. Information - whether from a questionnaire, a written report, an interview, a work sample, or any other source - is of high quality if it is simultaneously reliable, valid, and useful. By reliable I mean error free. By valid I mean that the meaning of the information is known and, at the same time, is what we intend to use for the kind of evaluation we are undertaking. And I mean useful in two broad senses, both in the sense that the information serves its purpose - e.g., helps improve faculty performance - and in the sense that it is cost/effective, in the definition of cost which includes not only dollars and cents but less tangible costs like faculty and student morale and institutional image.

We can evaluate the reliability of our evaluative information in at least three ways, the standard test-retest and internal consistency paradigms, and a third paradigm upon which I tend to put considerable emphasis, transferability.

The test-retest paradigm can discover unreliability in the sense that data gathered on two (or more) different occasions differ, given an unchanged subject of the data. For example, if a student ratings questionnaire is given at two different times (same students, same unchanged instructor) and the ratings are different, then to the extent of that difference the information is unreliable. Unfortunately it's extremely difficult to know which set of data, the first or the second, is the better reflection of the true situation. Without an experimental study, all we can really tell is that there is a difference where there should not be. (I'd like to interject here that simply giving a ratings questionnaire to a class during the fifth and eighth weeks of a term is not sufficient; we need to make sure that all relevant variables are under control, e.g. that the instructor who is being rated has not changed during the intervening period. The only design I've been able to think of is to play the same television tape on two occasions and have the same students rate the instructor each time. If ratings of this instructor are different on the two occasions, there is reason to doubt the reliability of those ratings.)

The second standard way to study the reliability of information is to examine the data to see if each respondent was consistent when he should have been consistent. For example, if a student ratings questionnaire contains a number of very similar questions about a specific instructor trait, like organization, and a student's response to those questions is highly variable, sometimes high, sometimes low, we might distrust his answers. Of course we have to be sure that there is no legitimate reason for this variability - that, for instance, the variability does not indicate that the instructor's lectures were organized, but his answers to questions were disorganized. One might check this to some extent by comparing one student's pattern of responses to these organization items to the average pattern of his classmates, or the pattern of each individual classmate. If everyone shows the same pattern of variability, the reliability is more likely to be legitimate. Fortunately, a good statistician's standard bag of tools can provide this kind of information quite readily.

The aspect of reliability that intrigues me most is what Cattell calls transferability: information about the same thing should say the same thing, no matter from whom it comes. To use another example from teacher rating, if different sources of data disagree - either across sources, as when students' teacher ratings and their instructor's self-ratings disagree, or within sources, as when students disagree among themselves - then I think we have prima facie evidence of unreliability. Again, it's hard to know which of the sources of data is the more "correct"; to find this out would require an experimental design with an adequate external criterion. It might well be that such differences are legitimate, but until the legitimacy has been demonstrated the fact of disagreement should raise a flag cautioning possible unreliability.

What is intriguing about the concept of unreliability is its implication for what we usually call "correlates of data," e.g., correlates of student ratings. While this correlational information is important and useful in itself, I think it becomes still more useful when we look at the associated rater variables - like year in school, IQ, sex - as indicators of levels of variables over which ratings, in order to be reliable, must remain the same. Thus, an instructor's rating is reliable (in this sense) if students of various years in school give him the same rating, and if students of various levels of intelligence agree. Again, there can certainly be legitimate reasons for difference in ratings across levels of these variables, but my point is that these differences need to be studied. For example, suppose for whatever reason female students tended to give their instructors more generous ratings on certain items. Consider then the case of two hypothetically identical instructors, one whose class is composed of all men, the other's of all women. The latter instructor could be rated more favorably simply because of this "sex effect". And this phenomenon is not restricted to student variables. A similar situation exists for situational variables like class size, hour of the day, and whether or not the course was required of the students.

One can, however, control these effects either at the item-selection stage of questionnaire development by eliminating items which show such effects, or at the data analysis stage by statistically correcting for the effect, or at the data reporting stage by norming according to these effects. (To the response that eliminating items on this basis risks throwing away important information, I go back to the purpose of the evaluation and suggest that if the data were being used to develop a theory of instruction, such inconsistency would be relevant, and would have to be accounted for, but if the data are being used to make a decision about the instructor, these differences are probably a form of unreliability that should be eliminated). Fortunately, we have found it quite possible to develop a broad-spectrum instructor rating scale even after sex-linked items have been eliminated from the initial pool.

To conclude this discussion on reliability, I'd like to propose an ethic: that the required level of reliability varies with the purpose of the evaluation, some uses of the data demanding a substantially greater freedom from error than others. My own leaning is that evaluation for personnel decisions demands the greatest reliability, since the effects of error here are, in my opinion, more severe than for any other use of evaluation data.

Validity is the second important quality of information: Do the data mean what we think they mean, and is that meaning appropriate for the use to which we want to put the data? Validation can be of at least three types. Some degree of meaning can be attributed to data - again, either data from questionnaires or interviews, or whatever - by a relatively simple inspection of these data. For example, if knowledgeable people - experts - agree, on the basis of their total professional experience, that items on a ratings questionnaire do measure consequential aspects of teaching behavior, then ratings from that questionnaire take on some meaning. (Of course, there's the question of the reliability and validity of these experts' opinions, but that's another matter.)

An external criterion can add still more meaning. If student ratings relate to the frequency with which students elect further courses from an instructor, certain further meaning is attached to the ratings. If how much students learn (not necessarily the grades they get) relates to the ratings they give, a great deal of important information is added. Better yet, perhaps, if patterns of relationships are found between various external criteria and various different ratings items, more meaning still is supplied. By that last point, I mean that a considerable degree of meaning would be attached to ratings if it could be demonstrated that, say, student ratings of the popularity of an instructor would relate more highly to an external (preferably objective) measure of popularity than to indices, say, of learning; that ratings of teaching skills would relate more to objective learning criteria than to indices of popularity, and so forth through a series of logical and pedagogically acceptable hypotheses.

That line of thought leads to the third and final aspect of validation, one which is especially useful in the common case in which no external criterion is really adequate. From this process of construct validation, meaning is attributed to the data on the basis of information from a well articulated interlocking network of logical and empirical demonstrations of meaning. In other words, the process entails setting forth the total accumulation of known fact about the data - everything we know from research on faculty evaluation - in a logical "If-Then" framework. To the extent that "sensible" patterns emerge, the data become meaningful, the hypothetical construct "effective faculty performance" takes shape. To the extent that new hypotheses suggested by the framework are confirmed, the data take on still more meaning. And to the extent that facts conflict, then either our research or our logic is suspect and the meaning of the data is encumbered. The articulation of such a framework concerning faculty evaluation, I'm afraid, is still rather far in the future.

Just as an ethical principle rises from the notion of reliability, so too one comes from the idea of validity. Again, and for the same reasons, I would propose that the level of validity required of evaluation data varies with the purpose of the evaluation, and that data for personnel decisions require the greatest degree of validity. But data need to be not only reliable and valid; they need to be useful. "Useful" is a very broad word in this context. It means first that faculty evaluation information needs to work, needs to contribute (at least potentially - that is, if people choose to use it) to the improvement of faculty performance. Student ratings done to help improve teaching, for example, need to be able to help improve teaching.

In a still broader sense, data need to be useful in cost/effectiveness terms. Clearly, we need to consider the dollars and cents aspects of any system of evaluation. The computer-terminal system that I described earlier for providing information to help students choose courses would probably not meet common cost/effectiveness criteria. But the intangible cost of any data and of any system must be studied too. What does it cost in terms of class time to gather data? Is this time well spent? Is there a cost to our evaluation in faculty or student morale? Is the image of the institution helped or hindered (in the eyes of the public, including legislators and trustees, as well as in the eyes of faculty, students, and administrators)? All these kinds of questions come under the heading of cost of a system, and therefore, utility of a system.

So, in order to be able to say we have data - or a system - of high quality, we need to demonstrate the reliability, validity, and utility of the data.

Related to reliability, validity, and utility are certain considerations that moderate or qualify the data. Three prominent modifiers are responsibility, competency, and motivation. For example, a faculty member might receive an unfavorable evaluation with regard to the text he uses in his teaching or the apparatus he uses in his research. But if all the texts in his area are poor, or if the good texts are prohibitively expensive, or if the proper apparatus is not available to him, and if he is aware of all this, then he cannot be held so responsible for these deficiencies as the person who simply isn't able to distinguish good materials from bad. In the same view, the junior faculty member who is required to teach material with which he is not familiar and does a poor job is not so responsible as his senior colleague who chooses to teach the same course and teaches it equally poorly.

The competency of the sources of data to evaluate is another moderator, whether it's the competency of colleagues who have never set foot in a teacher's classroom to evaluate that teacher's teaching, the competency of students to evaluate the long-term effects of the instructor, or the competency of a chairman whose specialty is vastly different from a researcher's to evaluate that research.

And there is the question of motivation. We find certain phenomena in ratings of all kinds - rating too leniently, or rating too harshly, for example. But these same kinds of phenomena can affect all kinds of (subjective) evaluation data. There can be vested interests or psychological reactions of all sorts that usually will manifest themselves as "leniency effects" or "stringency effects". Some kinds of statistical machinations can reduce some of these effects, but it's unlikely that statistics will ever control all of them. Consequently any evaluation needs to consider what these moderators can do to the reliability, validity, and utility of the data.

Sources of Measurable Data

Where do these data that I've been talking about come from? The possible sources of information about faculty performance are relatively obvious: students (present or previous), colleagues, administrators, members of the community, specialists in relevant fields, and the faculty member himself. From each of these sources we can get subjective information - opinions - about at least some aspect of faculty performance. From students the information we can get might be either subjective - like ratings - or objective - like the performance scores I've stressed. (It is conceivable that we might some day be able to get objective information from the faculty member himself, if there were, for example, a reliable and valid "How well do I teach" test; but to my knowledge no such test exists today.)

(It is worth pausing here to dispel too common a misconception about the "objectiveness" of ratings. The fact that questions are couched in "objective-looking" multiple-choice phrasing and can be processed by a computer doesn't in any way alter the fact that all ratings are subjective personal opinions. Now, we can certainly decide - and there is nothing wrong with this so long as we are aware of what we're doing - that opinions are the data we want for whatever the purpose of our evaluation: in that case, all we really need to do is demonstrate the reliability of the ratings. If, however, we want something other than opinion upon which to base our evaluation, then we need to relate the ratings to some external and more objective performance criterion: a learning criterion, perhaps, for evaluating teaching, a "correct outcome" criterion for research, and so forth, to the extent that criteria can be discovered.)

When I list these various sources of evaluation information, I do not mean to imply that these different types of people are all equal in the quality of information they can give about any aspect of faculty performance; neither do I mean to suggest a preference for anyone over another. But I would be extremely interested in seeing a well designed transferability study for the evaluation of teaching in which, say, ratings from students, colleagues, administrators, and present and former students were all compared to one another first in terms of their reliability and second -- more important -- in terms of their relationships to an external performance criterion (student learning). The reason I emphasize this point is that it is entirely too easy to approach student ratings with the stringent set of data-quality criteria that I've outlined, and simply to "badmouth" student ratings. It's entirely too easy to criticize student ratings in absolute terms without paying any attention to the quality of student ratings relative to each of the other kinds of evaluative information available to us.

But my intent here is not really to hold a brief for any one source of data over another -- only to say that no system of faculty evaluation can claim to be complete unless it has seriously studied the data from each of these sources. I doubt that any of these sources can be safely neglected in the research and development of a system of faculty evaluation, because I imagine that we will find one source most helpful for the evaluation of some kinds of faculty activity, other sources better for other kinds.

In this regard, it's interesting to look within each source of data and ask if certain students, certain colleagues, certain administrators, and so forth, might furnish more reliable, valid, and useful data than their peers. It would be a relatively simple matter to manipulate existing data to discover subsets, say, of students whose opinions more than their classmates' relate to a learning criterion. Identifying these students in terms of various personality and demographic variables could be informative indeed. It might even provide a way of sampling just certain opinions from future evaluations, those whose judgments are probably more sound than their confreres.

Media for Gathering Data

The media that are available for gathering and reporting data are another consideration. Pencil and paper still seem to be the quickest way to provide information; the questionnaire is inexpensive to provide and to analyze. But questionnaires are not necessarily the most efficient (cost/effective) means of garnering information. This is pure speculation, but it's possible that evaluations for improving teaching might be better served by some other medium--e.g., audio--or video-tapes.

I make this allusion to tapes because I suspect that they can provide some more meaningful kinds of information than the usual questionnaire. What makes me uncomfortable about questionnaires is the usual way in which data are reported. Frequency distributions, means, standard deviations, and deciles can certainly summarize a great amount of information, but these statistics have their drawbacks. I'm not really referring to the fact that many faculty don't understand statistics or are repelled by them; faculty are educable. I'm more concerned about the faculty member who receives low ratings: what can he do to change? To tell me that I am disorganized is not necessarily to tell me how to become better organized, and that fact makes me wonder how responsible--as well as how sensitive--we're being when we simply run ratings through computers and provide routine statistical analysis. I would be most pleased to have access to a Faculty Counseling Bureau where experts in the various arenas of faculty behavior could provide reliable, valid, and useful guidance to faculty who are trying to improve their performance. Some schools apparently have facilities of this sort. Ours has no such formal structure (although there are some informal avenues open--e.g., colleagues who are willing to share their experience and offer suggestions) so we have been experimenting with using the computer to generate prose narratives that expand on the basic data of teacher ratings. The computer examines an instructor's ratings profile and prints out personalized sentences that offer suggestions for changing low ratings and that reinforce high ones. But the computer approach and the Faculty Counseling Bureau approach share one major weakness: How can we know that the suggestions we offer are reliable, valid, and useful? At this point in time, until more research is in, all we can do is try to be reasonable, and acknowledge publicly that our counseling is highly subjective.

Beyond tapes, there is personal verbal communication--talking.

As awkward and frustrating as it might be, as much diplomacy as it might sometimes require--there is still no substitute for face-to-face communication. While data-quality problems--reliability, validity, utility--are extremely hard to deal with in verbal communication, the clarification and amplification of meaning and the exchange of views that can be accomplished through speech cannot be surpassed by any other medium. Some of our faculty have been urging a combined approach to self-improvement evaluations in which personal exchanges between students and teachers supplement the information gathered by ratings forms. I know of no data to support the utility of this approach, but the idea is most reasonable and the reports from people who have tried it have been good.

Temporal Considerations

A time dimension needs to be considered with regard to when information is collected and used and how it is reported. I do not want to bring up the issue of whether student ratings should be gathered before or after exams; this is largely an empirical question for which I have no data. Nor do I want to dwell on the dangers of all instructors asking for ratings during a single week so that students might be asked to fill out four or five questionnaires that many of them consider noxious or inane. This is essentially a question of student motivation which I think can be best met by public demonstrations that student responses are valuable, that someone pays attention to the ratings and that something happens because of them. It can also be helped by convincing faculty to use ratings sometime before the last weeks of the term so that, first, ratings aren't deemphasized by impending exams and second, so that there is at least a chance that the information a group of students provides might be of some direct benefit to those particular students. In short, we need to minimize the aversiveness and maximize the reinforcement to the respondents.

But the developmental issue I really want to emphasize is that almost any kind of faculty evaluation system attempts to measure typical performance as distinguished from maximum performance. We are therefore sampling behaviors, and our data might--should, in fact--reflect the whole range of behavior variation. An instructor might have a great day or a lousy one, and ratings will reflect that. He might have a great quarter or a lousy one, and ratings will reflect that. I think we need to file evaluation data term by term so that developmental patterns of evaluations can be studied. Some faculty and some departments routinely store such information for this very purpose. It's also feasible, in situations where ratings are centrally processed, to include in the instructor's print-out summaries of past ratings for comparison with current ones. The point is that faculty performance ought to be examined developmentally, not just at one point in time.

Let me make one last remark in this respect--one concerning the transferability of data. If we choose to look at performance evaluation longitudinally, we need to be sure that the data are transferable. That is, since a different class of students is presumably involved each term, we need to be sure that any differences (or similarities) across terms is due to differences (or similarities) in the instructor's behavior, and not due to the changing group of students.

Consequences of Evaluation

I have mentioned various groups of people--students, faculty administrators, the public, and the faculty member himself--as sources of data. We need also to look at them in terms of the consequences they might enjoy or suffer as a result of the evaluation. Favorable or unfavorable evaluations might have good and/or bad outcomes, and these outcomes might affect people in any or all of a number of ways.

To give a few examples:

An uncomplimentary evaluation could certainly hurt the career development of a young untenured faculty member. But it could also enhance his development if it contributed to some appropriate behavior change or if it guided--or forced--him into circumstances in which he was more likely to be both satisfied and satisfactory.

A complimentary evaluation, on the other hand, could clearly help confirm or improve a faculty member's status (and remuneration). But that same good evaluation could also excite the envy of his colleagues, which could ultimately be more harmful to him than a bad evaluation might have been. A good evaluation, paradoxically, could lead a chairman to "urge" a person, say, who loved research but who happened to be a good teacher to increase his teaching load at the expense of time for research. Or vice versa.

Consider the chairman of a department. Any kind of evaluation may well raise problems for him--especially evaluations for personnel decisions--because any differential treatment of faculty may damage morale. Unfavorable evaluations of any of his faculty must be especially troublesome for the chairman--more so than for the higher-level administrator--because the chairman is most likely the person with the immediate responsibility for painful decisions (e.g., firing a colleague or refusing him a pay increase) or even for pointing out deficiencies. (Some chairmen, though, I'm told, have learned to cope wonderfully well!)

The student, too, runs some risk--though apparently less than either faculty or administrators. (Perhaps this might have something to do with students being generally more in favor of evaluation than either faculty or administrators!) It's hard to see how a student could be endangered by providing a favorable evaluation (assuming the evaluation is of high quality). But it unfortunately does not strain the imagination to think of unpleasant consequences that might befall the students if evaluations they gave were highly uncomplimentary. Hopefully this distasteful situation is less common than the emphasis on anonymity in ratings would lead us to believe. On the other hand, one would expect students to profit over the long run from any kind of high quality evaluation of teaching. For that matter, it would not be hard to build the argument that any faculty member--and the entire academic community--would profit over the long run from high quality faculty evaluations.

But enough about consequences. The human ego is of such complexity and creativity that no adequate listing of the possible consequences of evaluation seems possible. It's enough at this point simply to express the concern and to try to anticipate the most likely consequences.

Institutional Goals

The final set of considerations I want to discuss pertains to the goals of the institution, either as they pertain for the institution as a whole or for any part of the institution--division, department, program, or course. I would think that the goals of almost any institution would include at least some degree of teaching, advising, research, governance, and public service. If this is the case, or whatever the institutional goals might be, however general or specific, I think evaluations need to be considered in light of those goals. All I mean here is that in a school with primarily an instructional emphasis, it doesn't "count" so much that a faculty member may be an excellent researcher; he needs to be a good teacher. Conversely in a research institute, the faculty member's teaching is of less concern than his research. And in land-grant colleges the public service role is perhaps more prominent than in private colleges. I'm suggesting here the need for a correspondence between the institution's goals and it's members' behavior. But the other side of the coin is appropriate too: when a member's goals, manifested by his performance, are different from the institution's, both parties need to assess the legitimacy of their priorities. Thus a person in a small college who is a skilled researcher but a poor teacher might decide to move to a research institute (or a research job in a teaching school); or the school might decide that research is a more tenable goal than it had previously believed. Thus some major universities have reminded themselves of the place of teaching in the list of institutional priorities.

Conceptual Schema

We've spent a substantial amount of time talking about eight different kinds of considerations that deserve attention when we plan or study systems of faculty evaluation. I don't suggest that these eight encompass all the considerations there are, nor do I consider all eight equally important; but I do believe each merits attention.

In the time remaining, I'd like to try to build a conceptual schema that takes account of all these considerations. What I've really been trying to do this morning is not just discuss some random concerns about faculty evaluation, but to lay out in a more or less organized fashion these different considerations. Now I want to try to draw them together.

My basic tools are 2 or 3 dimensional figures--after the fashion of Cartesian coordinates.

Insert Figures 1 & 2 about here

Let's fill in the coordinates. On the X axis, we can list the different aspects of faculty performance, teaching, advising, research and publication, governance, and public service, each heading with all the specifications I've described.

Insert Figure 3 about here

On the Y axis we could add the sources of measurable data: students, colleagues, administrators, the public, the faculty member himself, and so forth.

Insert Figure 4 about here

Thus the upper left intersection refers to evaluation in which the students are the source of evaluative information about the different components of the faculty member's teaching performance and the descending cells concern student information concerning advising, research, and service.

We can add a third dimension: Quality of Data, or reliability, validity, utility, and moderators.

Insert Figure 5 about here

The top-left-and-front-most cell now becomes a consideration about the reliability of student information concerning teaching performance, and the descending cubes refer to the reliability of the information students can provide about the other arenas of faculty activity. So we're dealing with a three-dimensional figure that can describe X times Y times Z specific considerations about faculty evaluation, where each cell represents a "consideration". (There are other things we can do with these cells, as we'll see shortly.)

Now I'm going to break the laws of physics and go into the fourth dimension, the purposes of the evaluation.

Insert Figure 6 about here

The top red cell talks about the validity of student information about teaching for the purpose of improving that teaching, the middle red cell about the validity of that same information for personnel decision, and the bottom red cell about the validity of information from students for helping other students select courses.

The green cells down the X column in the top figure then talk about validity of student information for improving various faculty activities other than teaching: advising, research, and service.

The green columns along Y in the middle schema talk about the comparative validity for personnel decisions of information about teaching gathered from the faculty member himself, his colleagues, administrators, and so forth.

And the green column along Z in the bottom schema asks about the reliability, utility, and moderators related to student information about teaching intended to help other students select courses.

It would be possible to set all eight of our dimensions into a schema like this, but the model would be so unwieldly as to become useless. So let me just mention the remaining four considerations for review: the media for gathering and reporting information (questionnaires, audio and visual tapes, computers, and personal confrontations), the temporal component of evaluations (for longitudinal patterns of interpretation), the consequences of evaluation to each of the people involved, and the goals of the institution and the meaning these priorities add to or subtract from the evaluative data. Each one of the dimensions interacts with the four dimensions already presented in the schema. Since there doesn't seem to be a reasonable way to include these in the schema (although we could, at the cost of some of the interactions, substitute them one by one for Quality of Data on the Z axis), I'd at least want to see them included as footnotes to each cell in this model.

I need to make a few remarks about the flexibility of this model before going into a brief description of its applications. The model I've sketched is based on my own reflections about our institution, but any part of the model can be changed to fit another school. For example, the list of faculty behaviors--axis X--can be lengthened or shortened or in any other way modified. The Institutional Goals could be changed; so could any of the other components. (I would, however, hesitate to change the Purposes of Evaluation or the Qualities of Data, except perhaps to make them more specific.) In short, the thrust of this whole presentation is that we need to spell out these "considerations," the important components of faculty evaluation and then cast them into a schema such as this in order to see in detail the problems that we're facing. I'd hate to see this thrust hindered simply because a few of the parts of the model didn't apply to another school.

For the final few minutes I'd like to talk about uses for this schema. What's it good for? Because it's only a glorified outline, it can do whatever an outline can do. It can provide the structure for a talk (as it has today, to a large extent). Or it can guide a literature review or a research program, pointing out questions in each of the cells that need to be answered by work already done or yet to be done.

The model also seems to be a powerful tool for building or for criticizing different instruments and, better still, for developing or evaluating systems of faculty evaluation. For example, suppose we want to develop a student ratings questionnaire. Item writing can be guided by the first parts of the Faculty Behaviors dimension. Each of the cells on the X axis can hold any number of items that attempt to measure the particular aspect of faculty behavior. Because we want only student information at this point, we stay with the first column on Y. The item retention and validation phases of questionnaire development can progress (in whatever order) across each element on the Z axis--validity, reliability, utility and moderators. Further considerations about the items arise as we move across all the other dimensions.

The same approach, using different rows and columns, holds for the development or criticism of instruments for colleagues' evaluation of a faculty member's research, for self-evaluation of one's own committee work, or for any of the other X times Y possible instruments.

An analogous procedure can help us build a system of faculty evaluation. First we could determine with the help of the X axis which behaviors we want to evaluate. Then, with Y, we could select the sources of evaluative information that seem most appropriate for each behavior. We might consider how to gather the information by examining the media dimension. And we could move across Z (Data Quality) to evaluate this whole battery of data-gathering devices. Finally, each of the dimensions could help us by pointing out further considerations that our system needs to attend to.

I've found these to be the prime applications of this model--outlining my own thoughts about faculty evaluation, guiding me through the literature, directing our research program, and aiding in the development and/or criticism of instruments for any aspect of faculty evaluation. I want to stress the flexibility of the schema, its ability to tolerate more or fewer dimensions and the modification of any of those dimensions. And I want particularly to note the fact that the further each of the basic dimensions is specified--the more specific the listing of faculty behaviors, for example--the more complete the model is and the more it can help make order out of the chaotic field of faculty evaluation.

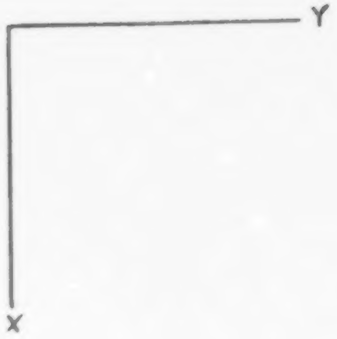


Figure 1:

A Two-Dimensional Figure

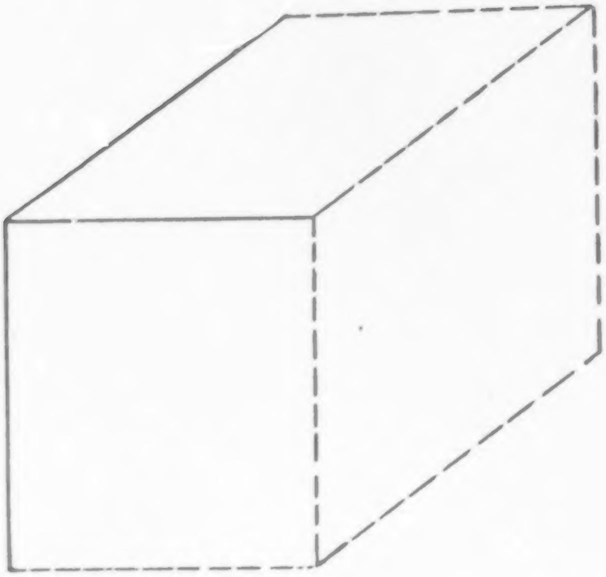
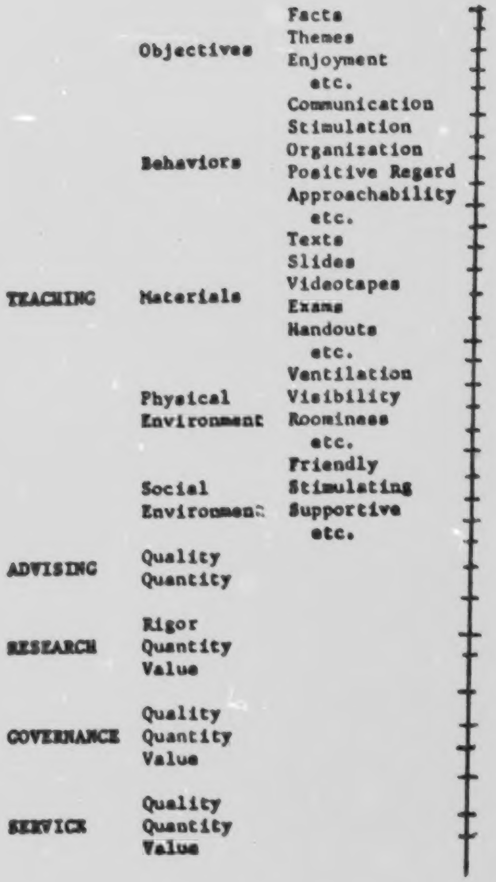


Figure 2:

A Three-Dimensional Figure

BEST COPY AVAILABLE



TEACHING	Objectives	Facts Themes Enjoyment etc.
	Behaviors	Communication Stimulation Organization Positive Regard Approachability etc.
	Materials	Texts Slides Videotapes Exams Handouts etc.
	Physical Environment	Ventilation Visibility Roominess etc.
	Social Environment	Friendly Stimulating Supportive etc.
ADVISING	Quality Quantity	
RESEARCH	Rigor Quantity Value	
GOVERNANCE	Quality Quantity Value	
SERVICE	Quality Quantity Value	

Figure 3: A Somewhat Expanded List of Faculty Activities

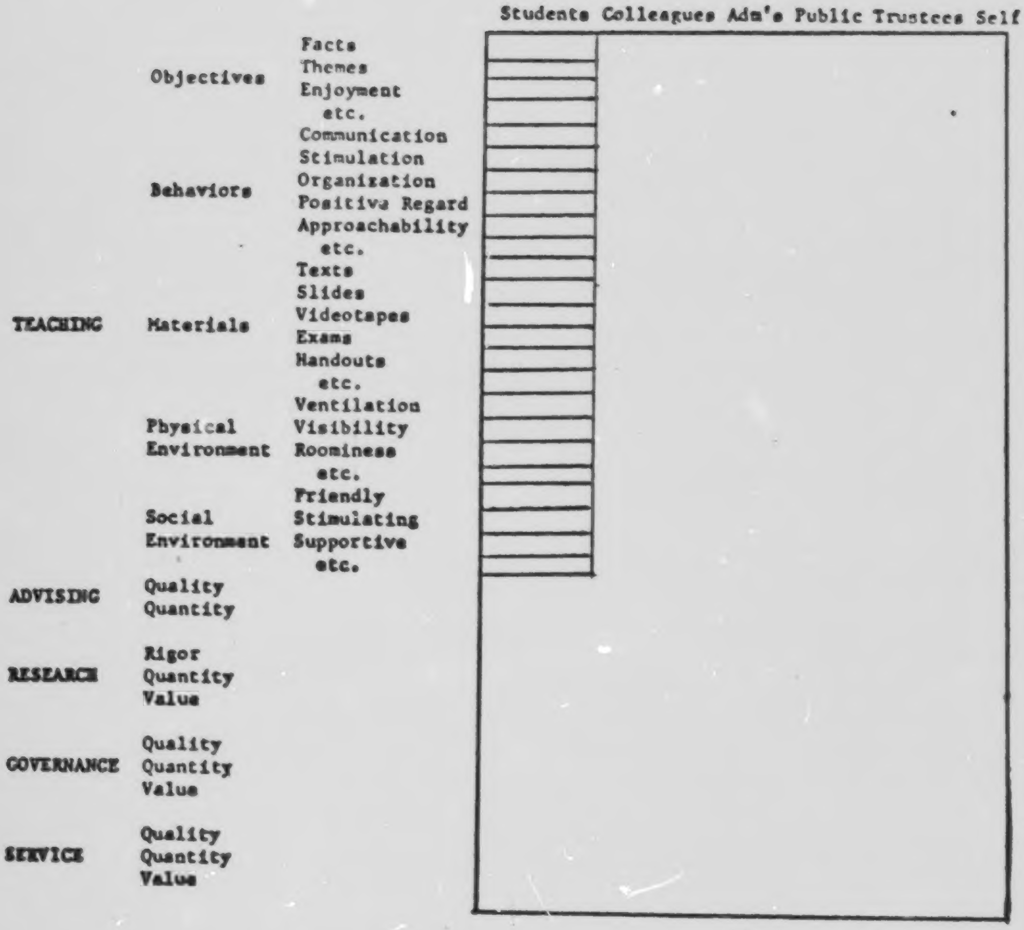


Figure 4: Schema in Two Dimensions

BEST COPY AVAILABLE

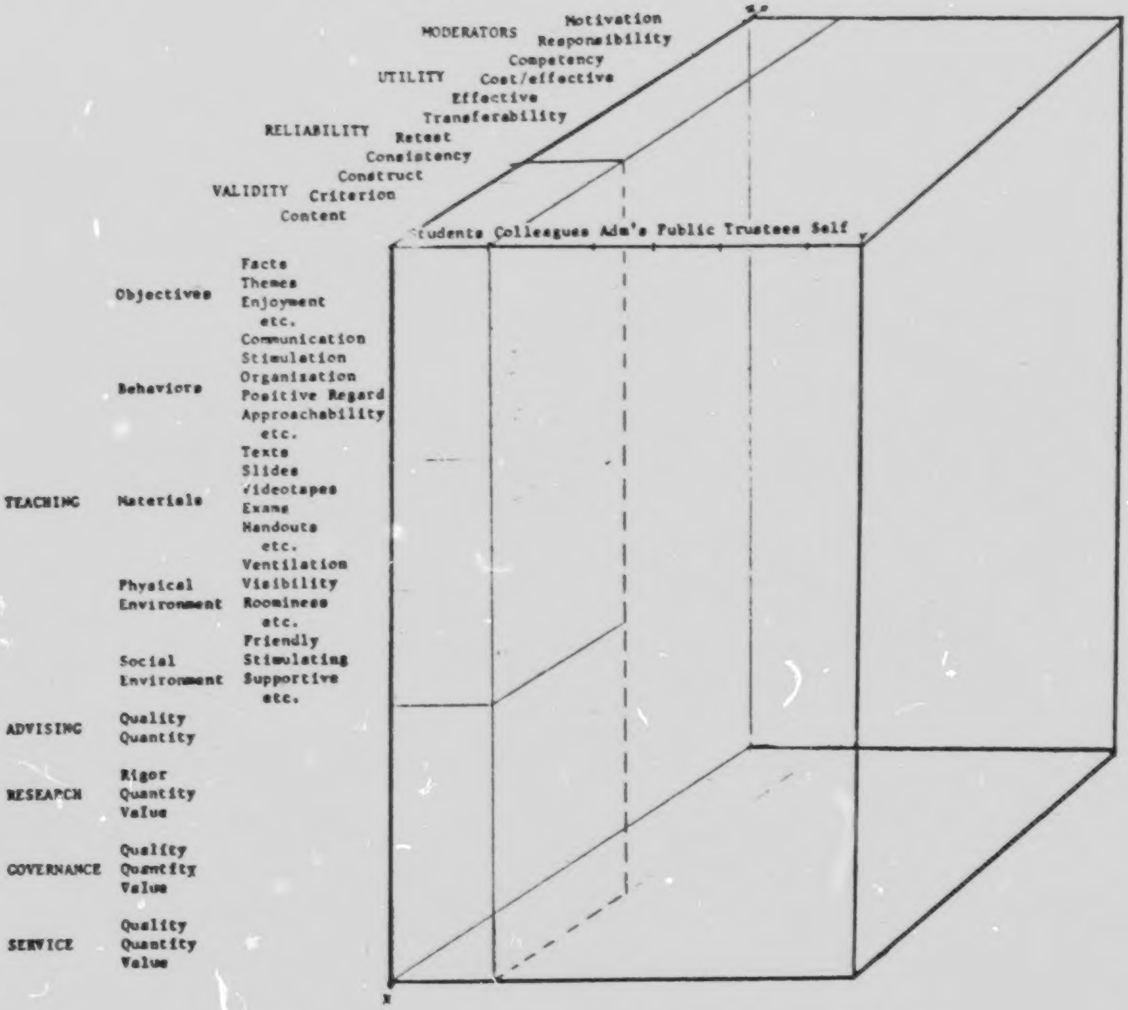


Figure 5: Schema in Three Dimensions

BEST COPY AVAILABLE

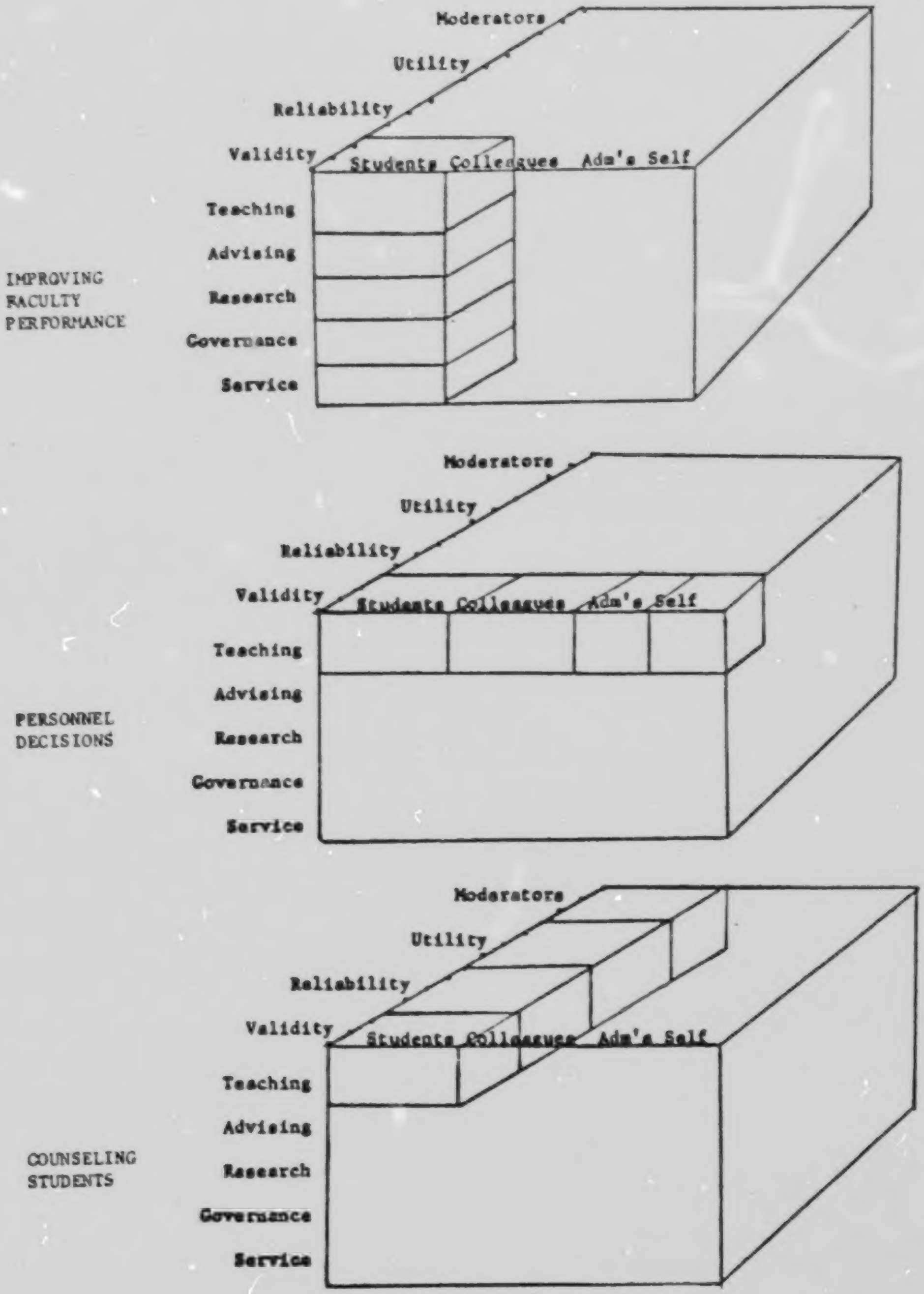


Figure 6: Schema in Four Dimensions

A SYSTEM FOR HELPING TEACHERS TO CHANGE THEIR AFFECTIVE BEHAVIOR THROUGH FEEDBACK

Bruce W. Tuckman

Rutgers University

I. Introduction

What I am going to do at the outset is present a general model of teacher behavior, and talk about some research findings that I obtained in my work with my students having to do with changing the behavior of teachers by giving them feedback. I am going to describe two studies that led me into this area of interest and led me to draw some conclusions upon which my operational approach is based. Then, based on these findings and some related theoretical notions, I am going to present some general rules that I see as descriptive of the change process in general and appropriate for changing the behavior of teachers in particular. And finally, I will describe a specific technique that I have developed and begun to use for providing feedback to teachers.

Let me emphasize at the beginning that I will be talking about feedback, not evaluation. Evaluation is a term that has many connotations, not all of which are part of its technical meaning. But I will be talking about feedback, about individuals gaining an awareness of their own behavior in order that they might change it in desired directions (with the emphasis on the word "awareness").

II. A Model of Teacher Behavior

Figure I illustrates a general model of teacher behavior which helps provide the context for an examination of the change process.

The focal element in this model is teacher behavior. This relatively simple model portrays (1) teacher training experiences, (2) the teaching environment in which the teacher is located, and (3) characteristics of the student that confront the teacher, as factors that cause the teacher to be what he or she is, a mixture of style, skills, attitudes, and so on. Now this person, the teacher - this collection of style, skills, attitudes, etc., then goes into a classroom and teaches, that is, produces teaching behavior. The result of that teaching behavior is student outcomes.

What I am particularly interested in is the connection between the teacher as a person and teaching behavior. There is (as you can see in Figure 1) a loop connecting the teacher as a person and her teaching behavior. This feedback loop must be based on awareness. In other words, if the teacher monitors her own teaching behavior and, by monitoring it becomes aware of what that behavior is and how it affects students, then as a result she can make modifications in her style, skills, attitudes, and so on. If this monitoring or awareness does not occur then the likelihood that changes will occur are minimal. Since the first three factors (training, environment, and students) are things over which the teacher does not typically have much control - the students come and go, the training is already done for the most part (except for limited in-service experiences), and the teaching environment is quite complex, the teacher as a person is going to be invariant unless a feedback loop between teaching behavior and the teacher exists. Creating that feedback loop, that awareness, is what this presentation is about.

III. Some Relevant Research Findings

How can teachers be made aware of how they are behaving in such a way that they can change their behavior in a desired direction? Let me describe the first of two experiments that I completed with my students and colleagues directed to this question. The first of these (Tuckman, Hyman, and McCall, 1969) employed Flanders Interaction Analysis categories (Flanders, 1965) as the feedback instrument. (These categories are shown in Figure 2).

The first thing to decide is if you are going to give teachers feedback in what form should it be given, that is what instrumentality should be employed. Obviously you can observe a teacher teaching and then sit down with the teacher and, in the course of a discussion, pass on feedback to the teacher. But we have an intuitive feeling (that has since been reinforced time and again) that in order to affect someone's behavior, feedback has to be definitive; it has to be concrete. Assigning numbers to categories seems to be the most definitive, concrete information that can be transmitted. Thus, we decided against some kind of anecdotal reporting, or rap session, or something like that, and in favor of some systematic set of categories about teacher behavior that we could report. We chose the Flanders System.

The Flanders System (shown in Figure 2) has a number of categories and breakdowns (or scores), the major ones being teacher talk - the number of statements the teacher makes, student talk - the number of statements the students make, indirect influence - how the teacher reacts to what students say or gets them to participate, and direct influence - the lecturing and authoritative behavior of the teacher.

The Flanders System is a model for information transmission but a model for motivation is still required. There has to be motivation for change to occur. Simply providing people with information may not be sufficient to motivate change; we might say that information is a necessary but not sufficient part of the change process - motivation is also needed. Motivation was introduced in this study using a prominent social psychological model called cognitive dissonance theory (Festinger, 1957). (The tenets of this model appear in Figure 3). Basically, cognitive dissonance theory postulates that people are motivated to have a high degree of internal consistency, to be consistent in their attitudes, behaviors, perceptions, and so on. A person who thinks of himself as a good citizen believes that he behaves in a way that is consistent with that perception. A person who thinks of himself as a good teacher and believes that a good teacher behaves in a certain way, is motivated to behave in a way which is consistent with those perceptions and expectations. Moreover, where inconsistency exists, the person is motivated to reduce it, and may do so by changing the attitudes involved, changing the perceptions involved, or changing the behaviors involved; in short, by changing some element of his psychological system. If a teacher thinks, for instance, that a good teacher should not talk much and discovers through feedback that he talks a lot, he will experience cognitive dissonance resulting from the inconsistency between the self-perception that he, as a good teacher, does not talk that much, and the feedback that he talks a lot.

What can he do in this circumstance? He can either decide that a good teacher really should talk a lot (i.e., alter his expectation or self-perception) or he can begin to talk less (i.e., alter his behavior). In either event, he will produce more consistency, which is, according to dissonance theory, what he is motivated to do. Similarly, if a teacher believes that the manner to influence students is not to be directive but to be non-directive (or indirective) and this teacher gets feedback that he is, in fact, directive, again there is inconsistency. The teacher will be motivated to reduce this inconsistency. How can he do this? He can change his idea of what a good teacher should do (or what he, as a "good" teacher, does) or he can change his teaching behavior. Thus, cognitive dissonance theory says that people strive for consistency, i.e., that they are motivated to be consistent. If you can (1) show people that they are inconsistent, and (2) constrain them to deal with that inconsistency so they cannot weasel out (because there is that tendency), then they will change some element of that inconsistency. This is dissonance theory.

We now have in the study what we believe to be the two sufficient conditions for change: motivation and information. The Tuckman, McCall, and Hyman (1969) study dealt with the variation of both motivation and information. A group of 24 teachers were given a form that corresponded to the Flanders categories and asked to estimate the percentage of time that they spent in each of the categories. This was a measure of self-perception. Then an observer used the Flanders categories to code the behavior of each teacher. We now had, on the one hand, what the teachers said they were doing and, on the other hand, what they were observed to be doing, and we could determine the degree of inconsistency. The 24 teachers were separated into two groups: those whose inconsistency was greatest, and those whose inconsistency was least. In one group teachers were doing what they thought or said they were doing, and in the other group they were not doing what they thought or said they were doing. And, of course, based on congnitive dissonance theory, the group with the greatest inconsistency was expected to have the greatest potential for change because they had the greatest motivation.

Each of these two dissonance groups was divided into four smaller groups, each of which had one of the following experiences. Teachers in one of these groups were given the reports of the observer who went over each category and gave them verbal feedback. (E.g., "You say you are talking 70% of the time but the data show you are talking 90% of the time. You say students are talking 30% of the time but the data show they are talking 5% of the time.") Teachers in the second group were taught to use the Flanders interaction analysis coding system and then had to code one another. This was done with the expectation that the best way to give people feedback would be to give them a mechanism for self-feedback. Presumably, a teacher who knew the coding system would be giving himself feedback all the time - resulting in a powerful effect. Teachers in the third condition listened to tape recordings of their own classes; and thus had to develop their own feedback and self-assessments. Since listening to tape recordings is kind of the antithesis of concrete, definitive feedback, it was expected to produce little effect. (Somebody could listen to a tape recording of himself and get no feedback whatever). Teachers in the fourth condition were given no access to any information of any sort. The fourth group is what the research designer calls a control group.

When we looked at the findings in this study (shown in Figure 4), we found there were tendencies for the discrepancy between teachers' self-preception and behavior to become smaller; as we had predicted the group with the greatest discrepancy had the greatest tendency to reduce it. In other words, teachers who had the most motivation to change were the ones that changed the most. That seems to make reasonably good sense. We also found that in the verbal feedback condition (i.e., when we sat down and gave teachers exact, quantitative feedback) those teachers changed; they reduced their discrepancy the most. The teachers that were taught the coding system did not change. In retrospect, were I to do the study again, I would have those teachers who learned the system code themselves from their own tape recordings rather than having them code their colleagues as they did. Apparently, even though teachers know a behavior coding system they do not necessarily use it on themselves unless they are put in a concrete situation where they have to -- which is what we did not do.

It turned out in this study that the motivation factor, i.e., the size of the discrepancy, primarily affected teachers' self-perceptions. Teachers who had the greatest discrepancy were the ones who changed their perceptions most. In a sense you might say that that is a kind of cop-out. If a teachers thinks that a good teacher should not talk very much and then finds out that he talks a lot, he then decides to say that he really does talk a lot but that is O.K. This was the nature of the finding. It did not occur for teacher talk, only for what is called the indirect ratio. The indirect ratio is a very complicated notion containing many elements all dealing with how the teacher reacts to students. Because of its complexity, it is not surprising that teachers did not change their behavior on the indirect ratio; rather, they dealt with discrepancies on it by changing their self-perceptions.

Our primary interest here is not in changing self-perception, however, but in changing behavior. Teachers did significantly change their behavior as a result of the verbal feedback. They changed their behavior in terms of the one element in the coding system over which they had the greatest degree of control: the amount of their talking in the classroom. They actually talked less. There was a very strong tendency for teachers to underevaluate their amount of talking; in other words, teachers typically believed that they talked less than they actually did. The resultant discrepancy motivated teachers in the verbal feedback condition (who were aware of the discrepancy) to talk less. The size of the discrepancy did not seem to matter; it happened pretty much across both discrepancy groups. Thus, teachers in this experiment talked less as a result of feedback.

At this point we began to think we had something, so we started out again in a somewhat different direction. In the second of the two experiments (Tuckman and Oliver, 1968), we used a different strategy. We had, in the first experiment, looked at motivation and the particular kind of information, i.e., the Flanders system as judged by trained observers, as change factors. In the second experiment, we looked at student judgements on a student opinion questionnaire as the source of feedback, and instead of looking at motivation per se, we looked at the source of feedback as a change factor. Who did the feedback come from? Two sources were investigated, each alone and in combination. One of the sources was the teacher's supervisor (in most cases assistant principals) and the other source was the teacher's students.

Four groups were used: one group got feedback from students only; one group got feedback from supervisors only; one group got separate sets of feedback from both students and supervisors; and one group got no feedback (the control group). Feedback was given on an instrument called the Student Opinion Questionnaire (SOQ) developed by Bryan (1963) and shown in Figure 5. It is an instrument usually filled out by students but actually anybody can use it, supervisors or students. Feedback was the same in all cases, so everybody was getting the same. The initial judgments of students and supervisors on a particular teacher did not differ so that regardless of what source a teacher was getting feedback from, the feedback was essentially the same.

The feedback had to do with the following ten areas (see Figure 5): the knowledge the teacher has of the subject taught, his ability to explain clearly, his fairness, his maintenance of discipline, his understanding, how much you are learning, "interestingness" of the class efficiency and businesslike manner of the teacher, skill in making students think for themselves, and the teachers' general, all-around teaching ability. These are global kinds of judgments but they still give you numbers. And the numbers were put on a graph, and the teachers given a profile of how they were seen in each instance. Incidentally, the SOQ has some open-ended questions on the reverse side which were not used in the analysis but made available to the teachers. These items provide a place to write in what you especially like about the teacher, how you think the teacher should improve, what you especially like about the course, and how you think the course could be improved. No attempt was made to quantify this information. Teachers were further separated into three groups based on how long they had been teaching.

Now for the findings (which are shown in Figure 6). First of all, the feedback changes were in the negative direction in every case. What that means is not that the teachers were necessarily worsening over time; since it occurred in every condition, you have to conclude that something else was going on. What seemed to be going on is that if you asked students to judge their teachers in February and you asked those same students to judge their teachers in June, they would be less positive about those teachers in June. Perhaps spring captures their fancy, or possibly since the teacher is about to give them the grade, they see this as a chance for retribution. At any rate, it is an end-of-year effect. Essentially, the students become increasingly negative toward teachers in all conditions over time, but we still could evaluate the outcomes of the experiment in terms of which condition had the least tendency to become negative and which had the most tendency to become negative.

Th[illegible]nly feedback that had a positive effect, that is, that minimized the negative effect, was feedback from students. Feedback from supervisors, even though it was the same feedback as from students, moved the teachers in the opposite direction to that advocated by the feedback. If the supervisor said you are not fair enough, for example, the teacher became less fair. If the supervisor said you are not efficient enough, the teacher became less efficient. In each case the supervisory feedback caused the teacher to change in the opposite direction, whereas the student feedback was followed. In other words, the teachers changed in the direction advocated by student feedback and in the opposite direction to that advocated by supervisors (even though both gave identical feedback). And when you gave them feedback from both students and supervisors, the result was pretty much the same as from students alone. The supervisory feedback in this case has little effect.

This finding led me to become very concerned about the source of feedback per se, particularly since the supervisor plays an evaluation role, and evaluation, because of its personal, career relevance can be very threatening. However, I am not interested in evaluation; I am interested in feedback. I am interested in getting people to change based on their own inherent motivation. It is an internal process; feedback information is not for publication. The studies I have described were feedback studies, not evaluation studies. Teachers were assured that data would not be put into their files, i.e., that nobody would have access to the data. And yet, the data lead me to believe that the supervisor is viewed as an evaluator. Because it is very difficult for the teacher to separate the supervisor's role as an evaluator on the one hand, and as a source of non-threatening feedback on the other, it would be very difficult to make the supervisor part of the feedback process.

Data concerning the years of experience of a teacher were examined on the hunch that there might be a difference between more experienced and less experienced teachers vis-a vis their willingness to change. We did find effects that were not strong enough to be called significant but strong enough to bear repeating. The tendency we observed for the teachers with the least experience was to be most resistant to supervisor feedback, and the teachers with the most experience to be most resistant to student feedback. That is, the younger teachers were more receptive to student feedback, and the older teachers more receptive to supervisor feedback. This is interesting because the older teachers (older meaning eleven or more years of experience) were tenured and thus had no great threat associated with the supervisor. For these teachers, the supervisor was probably viewed as giving feedback and not doing an evaluation. This confirms my earlier supposition about what is happening when the supervisor reacts to the teacher.

IV. The Change Process: Rules of Effective Feedback

It may seem presumptuous to see someone attempt to explain the feedback process based on two experiments, but nevertheless I will try. Being in somewhat of a hurry to get out into the real world to actually see what kinds of changes can be produced (and having devoted much time to these experiments), I attempted to put together what I consider to be the twelve rules of effective feedback. These are shown in Figure 7. Let us consider each in turn.

The first of these twelve rules of effective feedback is that feedback must involve concrete behaviors or characteristics. If you want to talk about things that a teacher can understand and relate to, you have to make the feedback as concrete as you can. That is why numbers (i,e., quantifications) help. If you talk about this much of a quality now versus that much of the quality then, it becomes easier to communicate the information. Or, alternatively, you can say you think this much of the quality is good but you only have that much of the quality. You can bring the feedback to bear much more easily if it is concrete.

Secondly, the feedback must provide clear, incontrovertible evidence of exactly how you appear to behave. After it is given, if the teacher can say that that is your opinion, then it is not incontrovertible. It is important, therefore, to think in terms of a feedback system where the evidence is strong and compelling in order that it be accepted by teachers.

The third rule is that the feedback source must be reputable and believable and his or her intentions accepted. To a large extent this may eliminate supervisors. I do not think that teachers question their reputability and believability as much as they do their intentions. I think there may be a great limitation upon supervisors within the feedback process; this is not to say that supervisors cannot play a role in the feedback process but the issue of intentionality must be dealt with.

The fourth rule is that feedback must be in terms that the teacher can understand and relate to. One of the problems with the Flanders system is that the ratios, for example, are not easily understandable. Teachers cannot (as shown in the first study) behaviorally change these ratios. After all, who can keep in mind the three terms of the numerator and the one term of the denominator and change the three of the numerator up and the one of the denominator down all at the same time? It is just too complicated and thus not likely to happen.

The fifth rule is that the feedback recipient must have a clear ideal model of what his behavior or characteristics should be. If we are going to try to motivate teachers by creating some state of dissonance or discrepancy between the way they are perceived and the way they want to be, we must make sure that they are clear about the way they think they should behave.

The feedback recipient must also know what others expectations of him are (Rule 6). I think that that is an important factor in formulating a personal model about what kind of a teacher you want to be. In other words, what kind of a teacher you want to be must be based in part on what kind of a teacher people expect you to be. Obviously, you cannot behave in a way that meets students' judgments unless you know what they expect of you. If your peers are going to provide you with feedback, you must know their expectations. If your supervisors are going to provide you with feedback, then you must know their expectations.

The seventh rule is that you must make a commitment as to the way you would like to be. There must be a commitment in this system somewhere, otherwise you can weasel out. You must say at some point or another, "This is what I want to do. I don't want to talk so much. Talking so much is not good teaching." That is a commitment. It is like Weight-Watchers, or Smoke-Enders where your commitment is partly based on the money you pay. For $70 you might give up almost anything.

The eighth rule is that you must also make a public commitment to change (another similarity to the procedures used by Weight-Watchers and others). You cannot mumble this commitment under your breath so that nobody hears it, because if you do, it may be the kind of a commitment that you give up when the going gets rough. It must be public.

The ninth rule is that the feedback must create tension. That is, it must be dissonant with your self-perceptions or ideals and it must be internalized. This gets back to the idea of motivation. If you think something about yourself and you get feedback that confirms it, you will not change, and appropriately, you should not change - there is no tension. If you want people to change, you have to find out ways of giving them feedback that is inconsistent or dissonant with the way they see themselves, thereby creating the tension for change. That may mean having a fairly flexible feedback system - something to keep in mind.

The tenth rule is that the reception of feedback must not involve more than low risk, i.e., support should be provided. This is very important. Feedback, in any aspect of life - professional, avocational, etc., is not easy to accept. It is not easy for people to tell others how they see them and it is not easy to hear it, especially if it is not consistent with the way someone sees himself. Since feedback is something that does have a degree of inherent threat, one of the rules must be that at the same time you give feedback, you must provide some kind of support.

The eleventh rule is that models for change and for the support of change must be provided. A feedback system must be part of a model, that is, it must relate to other aspects of teaching behavior, and there must be the possibility to generalize from it. If a feedback system does not provide the possibility to generalize from it, the kinds of changes produced may be very finite and limited, as opposed to actually producing major changes in a teacher's teaching philosophy. In other words, a feedback system must deal with teaching philosophy.

And finally, the twelfth rule of feedback is that accountability (by now one of your favorite words) to your group must be maintained through continuing feedback. When I say accountability I mean accountability to the people who are providing you with the feedback. And in the model that I will advocate (later on), the people who will provide the feedback will be other teachers. In accountability, as I see it, you make a public commitment to your peers that you will attempt to accept and use their feedback, and in turn have accountability to provide them with feedback too. This is a kind of accountability that I believe can be lived with.. It is accountability based on the fact that you are asking people to help you, to contribute to your growth and development; therefore, you have a responsibility to give this feedback a serious try.

V. The Change Environment

Since feedback asan element for change occurs in a total environment, I would like to talk for a moment about what I call the change environment. The change environment is a critical component of change. Nothing will happen unless the environment has those characteristics that contribute to the change process. The components of the change environment that I identified are shown in Figure 8. The change environment, first of all, must have newness. If you are going to change, you obviously have to have something to change to; some "innovation." And if you have to change to be doing it, then that something for you is new. Be it accountability or behavioral objectives, they are new; feedback from peers is new; team teaching may be new for you or for your system; non-grading, differentiated staffing, and so on. All I am saying is that a critical element for change environment is having something to change to which will be new for the potential adoptor.

The change environment must also contain the element of compelling reality. This is unfortunately a "negative" aspect of the change environment, but it has to be present. This is the "shotgun." This is the father who comes rapping on the door of the young man who just left the hay loft with his daughter and delivers an ultimatum. That is compelling reality and it will motivate that young man to the alter in some short space of time. As negative as it might seem if you want to affect behavior, there has to be some kind of constraining or compelling reality. Threatening to burn down the school, for instance, is in some ways a very effective compelling reality. There is no way to get around the fact that that is going to get your blood flowing. It produces the kind of threat that does unfortunately seem to contribute to the change environment. If everything is nice, happy, pat-on-the-back, we-are-all-in-this-together, we-are-going-to-make-better-schools, then in my perception, nothing happens. Compellingness has to be produced by someone, be it the board of education, the superintendent, the principal, subgroup of teachers, the parents, or the students; someone has to hold a shotgun to the group that they are trying to change. That is the compelling reality.

The third element is called open participation, that is an honest opportunity to contribute to the change decision, and an honest willingness to be a part of the process of change. In the case of teacher feedback, this means saying: "I want to know how you perceived me and I am willing to tell you how I perceive you." And that kind of open participation represents a risk, and there is no way to finesse that point. It does not matter who you are; open participation is risky.

The first three elements - newness, compelling reality, and open participation, all represents risks of a sort and might be considered negative elements in some sense. On the more positive side are the last two elements both of which help us live with this risk and be willing to take it. The first of these is a problem focus, that is, the realization that what is called for is problem solving. We must realize that the tools and the skills of problem solving can be brought to bear in dealing with the problem that is prompting the change (and be allowed to use these tools and skills). In other words, we are rational people to some degree and our problems can be solved rationally. That focus is a critical part of the change process.

Finally, we have the element of support. Risk reduction and the maintenance of the entire system are dependent on what I call the group-and-leader. It may be an informal group like the wildcat strikers who are meeting in the basement of somebody's house to plan their next strategy or a board of education caucus, or it may be a formal group like the entire board of education or the teacher's union. At some point within the change process, there are groups that form and leadership that emerges, and these groups are a source of support. You can lean on them when things get rough. However, these same things can also be a oppositional force to the change process; they can provide the greatest resistance to change by using their support mechanism to avoid it. When that happens, the group is beginning to deny open participation. As soon as the group denys open participation, one of the elements of the model disappears and therefore change is not going to occur.

What I am saying is that you need all five elements of the change environment for change to occur. You can't have four of them, or three of them; you need all five of them. The newness, the innovation, produces the challenge (or you may call it threat). Compelling reality, the burning building, the subtle edict or whatever, provides the confrontation. Open participation allows for specific feedback. The problem focus creates a problem-solving-orientation, and the group-and-leader provide support for risk reduction. Taken together, these factors create in people a willingness to experiment, that is, to try things out, a willingness to show themselves, and a willingness to be receptive to others. And these are ultimately the major ingredients of learning and growth. This is perhaps a somewhat abstract and idealistic conception of change but it does provide a reasonable point of departure into the specific mechanisms for changing teacher behavior through feedback.

VI. The Tuckman Teacher Feedback System

Let us move on to the last step by putting together the data and intuitions from the two experiments along with some of the more general concepts that I evolved from them. Let us consider a feedback system that hopefully would become part of the larger educational system and help teachers to change their own behavior. I designed a form for this purpose which I called the Tuckman Teacher Feedback Form (or TTFF). (I figured that if Flanders could have a Flanders Interaction Analysis Form, then Tuckman could have a Tuckman Teacher Feedback Form. Certainly, nobody else was going to call their form the Tuckman Teacher Feedback Form).

The TTFF began as a rather long laundry list of adjectives each of which somehow seemed to describe a human element in behavior and each of which was paired with an opposite, e.g., original-conventional, passionate-controlled, impertinent-polite, patient-impatient, cold-warm, initiating-deferrent, and so on. I purposely tried to use adjectives that describe the human element in teaching. It seems that we have many other ways to evaluate, or provide feedback about, the curriculum, and there are many ways to provide feedback or accountability in terms of student performance. But after all is considered, the teacher as a human being still has the human element as a unique quality of teaching and most of the existing feedback or evaluation systems make no attempt to assess it. I am not presumptuous enough to think that I know how a teacher should be on these human elements (although, I think if we were to discuss it for even a few minutes we would reach a high degree of agreement). The point is that in my system every teacher has the right to specify what he thinks the good teacher should be, that is set his own goal, and work toward it. Moreover, the feedback is referenced only in terms of his own goal. So I am not imposing an arbitrary standard on teachers but attempting to introduce or reintroduce the human element back into teaching.

As I said, I began with this long laundry list of adjective pairs that I more-or-less picked out of the air. And when I thought that I had a long enough list, I recruited 80 of my students who were also teachers, administrators, or full-time graduate students at Rutgers and asked them to use these adjective pairs to rate one of their graduate instructors. I used a statistical procedure called factor analysis to analyze the data they provided. Factor analysis is a procedure that allows you to tell numerically or quantitatively when different things apparently mean the same thing to the same person. In other words, I can use the adjective "original" and "creative", and mean different things by them. Factor analysis can tell you the extent to which people in fact mean the same thing by these two terms by determining whether they use them in the same way when they are judging someone such as a teacher.

(The scale or form I came up with using this procedure, the TTFF, appears in the Appendix.) Remember that I made no attempt to be systematic in selecting these 75 adjective pairs to begin with; however, people just do not use 75 pieces of information to describe teaching behavior (or any kind of behavior for that matter). It is much too many. The factor analysis reduced the 75 adjective pairs to four factors. There were just four factors or clusters of meaning in this whole laundry list. Each is shown in Figure 9, and will be briefly described below.

The first factor I called creativity. The teacher who was creative was imaginative, experimenting, original, iconoclastic, uninhibited. adventurous, flexible and initiating, in constrast to the noncreative teacher who was routinized, exacting, cautious, conventional, ritualistic, inhibited, timid, dogmatic, and deferrent. Those pairs of words meant the same to the student judges (as evidenced by the factor loadings in the factor analysis), and I chose the term "creativity" as a way of trying to label what those words seemed to have in common. Thus, the student judges first reacted to the creativeness of a teacher, and seemed to do so in very personal terms.

I had a little more trouble naming the second factor. I called it dynamism. It seemed to me to be a combination of dominance and energy, and so I called it dynamism because I did not want to use a word that conveyed just energy. Dynamism has within it (according to the analysis) bouyant, extraverted, bubbly, and outspoken, all of which seem to refer to a teacher's energy level. This factor also includes aggressive, assertive, dominant, and direct which are dominance terms. It seems to mix together two qualities that I had thought were separable but were not distinguished by the judges.

The third factor I called organized demeanor; again I used a somewhat obtuse label rather than simply calling it organization. I did this because it includes more than just terms that refer to organization. True, it includes systematic, organized, purposeful, resourceful and knowledgeable on the one hand, but it also includes in control, sophisticated, observant, and conscientious on the other hand. It is more than organization; it is organization plus self-control.

Finally, the fourth factor I called warmth and acceptance. That describes the warm, sociable, amiable, patient, fair, gentle, accepting, thoughtful, polite teacher as opposed to the cold, unfriendly, hostile, impatient, unfair, harsh, critical, inconsiderate, inpertinent teacher. (Each of these factors could have included other words had I introduced other words into the laundry list.)

There is a specific scoring procedure for the TTFF based on a scoring form (which I have included in the Appendix). All of the items on a factor are not included in the scoring in order to avoid unnecessary redundancy. Also, the adjective pairs are written in both directions. (As you can see on the TTFF, some have their "positive" end on the left, some on the right.) This is just a good measurement strategy. If you put the positive adjectives on the left all the time and the negative ones on the right, someone can fall asleep and mark the left end on each and you turn out to be the greatest teacher in the world. Since no one wants falling asleep to be a factor, the items are written in both directions. As a result when the TTFF is scored, either the positive or the negative items must be turned around or scored separately. When that is done, a constant must be added so that the lowest score a person can get will be "1". This is done because it is much easier to deal with positive numbers than negative ones. The scoring procedure looks more complicated than it really is. The scoring form is quite explicit and, when used, makes the scoring process quite mechanical. Once the scoring is completed, the profile of the teacher on the four dimensions is plotted at the bottom of the scoring form so he can see how he has been judged. The resulting line between points provides the teacher with a basis to react to himself.

The steps in the total feedback system that I am proposing are shown in Figure 10 and described below. The first thing I would do in the feedback system is ask the teachers to fill out the TTFF describing "The Good Teacher." They may be describing themselves; I am not quite sure whether that matters. I am not willing to say at this point that the higher you are on these four dimensions the better a teacher you are. It may not be that simple. I would rather the criteria be what you yourself think a good teacher is, or what we agree consensually that a good teacher should be. Remember that nobody is being forced to change by this system so the more points of reference there are the better. Remember also that the basis for change is to be dissonance - dissonance between what you are and what you think you are. So I would begin by having a group of teachers fill it out on "The Good Teacher." Six or seven teachers within a school might be involved and each would be asked to fill it out on his own.

Then, the teachers would be given the opportunity to observe one another. This can be done by sitting in on one another's classrooms, or, if the facility exists, using closed-circuit television or video tape. Regardless of how it is done, the fact remains that you cannot judge a teacher's behavior unless you observe it, whatever the inconvenience. When a teacher is out of his own room you have to bring in a substitute. The process, as you will see, may also require some in-service time.

Then each teacher is given a consensus summary statement of ratings of him by the other teachers in the group, so he knows how his teaching behavior is perceived. At the same time, the teachers involved meet as a group to discuss the feedback. This is done so that the feedback is not conveyed by just an impersonal sheet that you find in your cubby-hole mailbox one day. It is not meant to work that way. The feedback is given in conjunction with group process.

In the next step, the teachers engage in what I call strength training (for want of a better term, and since somebody has already coined it that). Now that you see from the feedback form what your deficiencies are, you ask yourself what you can actually do in the classroom to overcome them. In strength training you learn how to create new strengths for yourself. You do this by discussing your deficiencies with one another and giving one another specific ideas about how to convert them into strengths. The teachers can even role play these new strength techniques on one another. At the same time, they try out these new strength techniques in their regular classes. Take dynamism, for example. If the teacher is not seen as being as energetic as he or she would like to be, the other teachers in the group might point out certain things about movement and modulation of voice and activity level that might make strengths out of these weaknesses. The teacher can then try these things out in her actual classes.

And finally, the teachers then observe one another a second time to provide a basis for determining whether there has been a change in behavior in the recommended direction.

VII. Summary

This paper has covered a lot of ground. It began with a model of teacher behavior that linked the teacher to his own behavior through awareness based on feedback. Two studies followed that showed that teachers would change their behavior based on feedback information telling them how they were perceived. These studies also indicated that dissonance between self-perceptions and the perceptions of others was a motivator of change, and that supervisors, traditional sources of "feedback" to teachers, had little effect.

Based on these studies, 12 rules of feedback were presented as a kind of operational philosophy of changing teacher behavior. These rules were further generalized to provide a conception of the change environment - those conditions that must exist for change to occur. Finally, the feedback rules and the change environment characteristics were incorporated into a total teacher feedback system (which I named after myself) which incorporated a feedback form and scoring system designed and analyzed for the purpose of providing teachers with the kind of information about themselves on which change could be based. The instrumentation was further nested in the group process to provide the mechanisms for change required by the change environment. The obvious next step is to try it out. This is now in process.

References

Bryan, R. C. Reactions to teachers by students, parents, and administrators. U. S. Office of Education, Cooperative Research Project No. 668. Kalamazoo: Western Michigan University, 1963.

Festinger, L. _A theory of cognitive dissonance_. Palo Alto: Stanford University Press, 1957.

Flanders, N. Teacher influence, pupil attitudes, and achievement. U.S. Office of Education, Cooperative Research Project No. 25040, 1965.

Tuckman, B. W., McCall, K. M., & Hyman, R. T. The modification of teacher behavior: Effects of dissonance and coded feedback. _American Educational Research Journal_, 1969, _6_, 607-619.

Tuckman, B. W. and Oliver, W. F. Effectiveness of feedback to teachers as a function of source. _Journal of Educational Psychology_, 1968, _59_, 297-301.

BEST COPY AVAILABLE

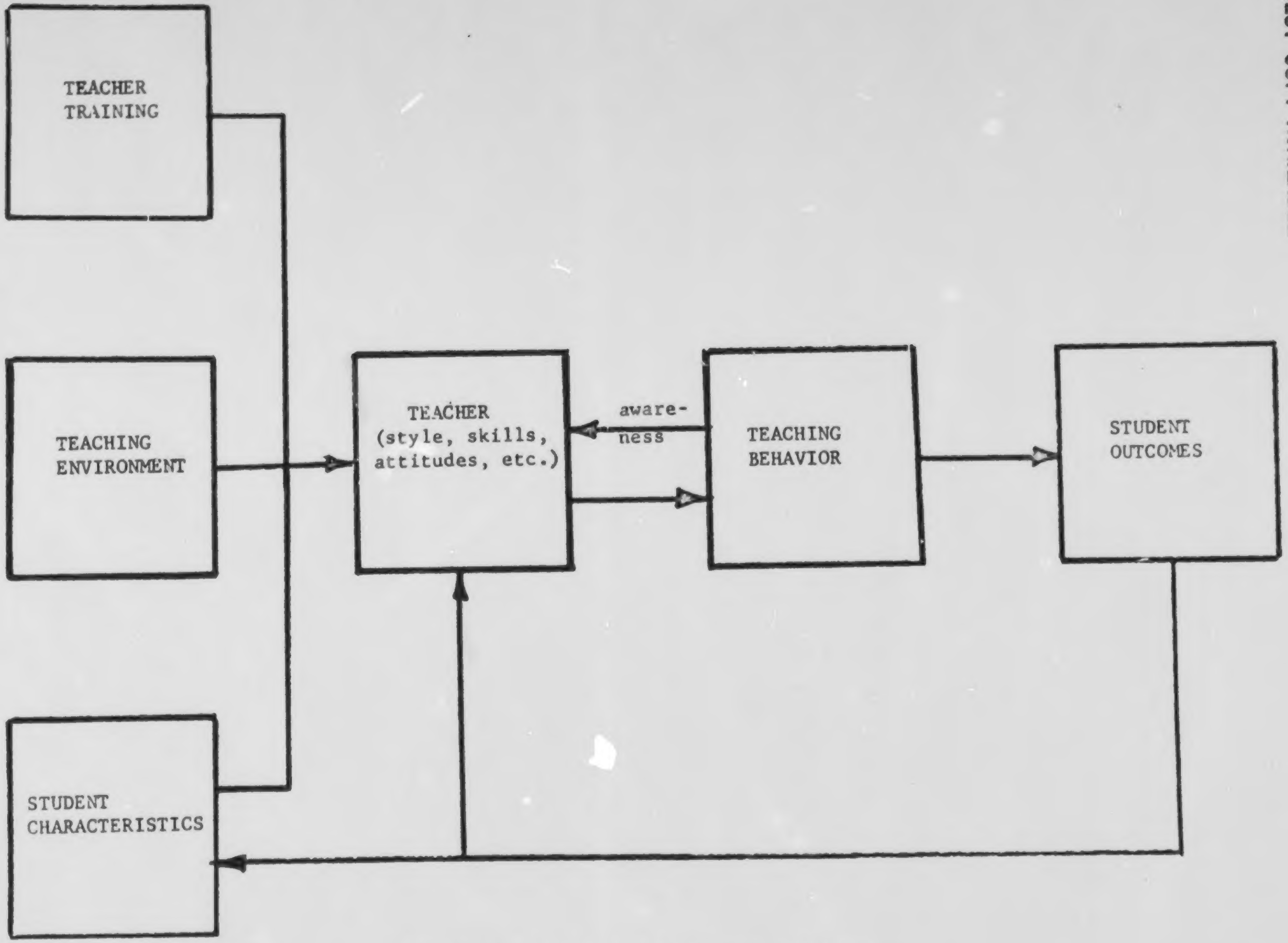


Figure 1. A model of teacher behavior.

TEACHER TALK	INDIRECT INFLUENCE	1. *ACCEPTS FEELING: accepts and clarifies feeling tone of the students in a non-threatening manner. Feelings may be positive or negative. Predicting and recalling feelings are included.
		2. *PRAISES OR ENCOURAGES: praises or encourages student action or behavior. Jokes that release tension, not at the expense of another individual, nodding head or saying "uhhuh?" or "go on"are included.
		3. *ACCEPTS OR USES IDEAS OF STUDENT: clarifying, building, or developing ideas or suggestions by a student. As teacher brings more of his own ideas into play, shift to category five.
		4. *ASKS QUESTIONS: asking a question about content or procedure with the intent that a student answer.
	DIRECT INFLUENCE	5. *LECTURES: giving facts or opinions about content or procedures; expressing his own idea; asking rhetorical questions.
		6. *GIVE DIRECTIONS: directions, commands, or orders with which a student is expected to comply.
		7. *CRITICIZES OR JUSTIFIES AUTHORITY: statements intended to change student behavior from non-acceptable pattern; bawling someone out; stating why the teacher is doing what he is doing, extreme self-reference.
STUDENT TALK		8. *STUDENT TALK-RESPONSE: talk by students in response to teacher. Teacher initiates the contact or solicits student statement.
		9. *STUDENT TALK-INITIATION: talk by students, which they initiate. If "calling on" student is only to indicate who may talk next, observer must decide whether student wanted to talk. If he did, use this category.
		10. *SILENCE OR CONFUSION: pauses, short periods of silence, and periods of confusion in which communication cannot be understood by the observer.

Figure 2. Summary of Categories for
Interaction Analysis
(Flanders, 1965, p. 20)

DISSONANCE THEORY

1. Beliefs about ourselves and our own behavior are potentially dissonant if we behave in ways that are discrepant from or opposite to the ways we believe we should or do behave. When we are made aware of this discrepancy (or consciously create it), dissonance is produced.
2. The amount of dissonance produced is a function of (A) the importance or centrality of the self-beliefs in question, (B) the extent to which the evidence of the discrepant behavior is incontrovertible and (C) the magnitude of the discrepancy between belief and behavior.
3. The presence of dissonance gives rise to pressures to reduce it (proportional to its amount) because it is unpleasant to experience.
4. Dissonance can be reduced by (A) changing our beliefs or perceptions of ourselves to bring them more in line with our behavior. (B) Changing our behavior to bring it more in line with our beliefs, (C) finding other evidence of our behavior which is more consistent with our beliefs, or (D) otherwise rationalizing or compartmentalizing the two so that nothing need change (such as negating the legitimacy and accuracy of the evidence about our behavior).
5. People can tolerate some degree of dissonance without changing but when the dissonance reaches a critical level, something must change.

Figure 3

Analysis of Mean Change Scores:

DISCREPANCY SCORE CHANGE (TOTAL)

	Verbal Feedback	Interaction Analysis	Tape Recording	Control		
High Discrepant	91	61	40	30	56	.01
Low Discrepant	33	-1	2	1	9	
	62	30	21	15		

.01

SELF-PERCEPTION CHANGE (INDIRECT RATIO)

High Discrepant	51	30	16	13	28	.05
Low Discrepant	15	-7	6	4	5	
	33	12	11	9		

BEHAVIOR CHANGE (TEACHER TALK)

	Verbal Feedback	Interaction Analysis	Tape Recording	Control	
High Discrepant	9	2	0	-5	2
Low Discrepant	4	-2	-4	-6	-2
	7	0	-2	-6	

.05

Figure 4. Changing Teacher Behavior Through Dissonance and Different Forms of Feedback. (From Tuckman, McCall, and Hyman, 1969.)

Figure 5. The Student-Opinion Questionnaire
(Bryan, 1963, p. 53).

Please answer the following questions honestly and frankly. Do not give your name. To encourage you to be frank, your regular teacher should be absent from the classroom while these questions are being answered. Neither your teacher nor anyone else at your school will ever see your answers.

The person who is temporarily in charge of your class will, during this period, collect all reports and seal them in an envelope addressed to Rutgers University. Your teacher will receive from the university a summary of the answers by the students in your class. The University will mail this summary to no one except your teacher unless requested to do so by your teacher.

After completing this report, sit quietly or study until all students have completed their reports. There should be no talking.

Underline your answer to each question on this page. Write your answers to questions 11 to 14 on the other side of this page.

WHAT IS YOUR OPINION CONCERNING:

1. THE KNOWLEDGE THIS TEACHER HAS OF THE SUBJECT TAUGHT?
(Has he a thorough knowledge and understanding of his teaching field?)
Below Average Average Good Very Good The Very Best

2. THE ABILITY OF THIS TEACHER TO EXPLAIN CLEARLY?
(Are assignments and explanations clear and definite?)
Below Average Average Good Very Good The Very Best

3. THIS TEACHER'S FAIRNESS IN DEALING WITH STUDENTS?
(Is he fair and impartial in treatment of all students?)
Below Average Average Good Very Good The Very Best

4. THE ABILITY OF THIS TEACHER TO MAINTAIN GOOD DISCIPLINE?
(Does he keep good control of the class without being harsh? Is he firm but fair?)
Below Average Average Good Very Good The Very Best

5. THE SYMPATHETIC UNDERSTANDING SHOWN BY THIS TEACHER?
(Is he patient, friendly, considerate, and helpful?)
Below Average Average Good Very Good The Very Best

6. HOW MUCH YOU ARE LEARNING IN THIS CLASS?
(Are you learning well and much? Are you really working?)
Below Average Average Good Very Good The Very Best

7. THE ABILITY THIS TEACHER HAS TO MAKE CLASSES INTERESTING?
(Does he show enthusiasm and a sense of humor? Does he vary teaching procedures?)
Below Average Average Good Very Good The Very Best

8. THE ABILITY OF THIS TEACHER TO GET THINGS DONE IN AN EFFICIENT AND BUSINESS-LIKE MANNER?
(Are plans well made? Is little time wasted?)
Below Average Average Good Very Good The Very Best

9. THE SKILL THIS TEACHER HAS TO GET STUDENTS TO THINK FOR THEMSELVES?
(Are students' ideas and opinions worth something in this class? Do students help decide how to solve problems and how to get their work done? Do they get at the real reasons why certain things happen?)
Below Average Average Good Very Good The Very Best

10. THE GENERAL (ALL-ROUND) TEACHING ABILITY OF THIS TEACHER?
(All factors considered, how close does this teacher come to your ideal?)
Below Average Average Good Very Good The Very Best

(over)

BEST COPY AVAILABLE

Figure 5. (Con't.)

11. PLEASE NAME ONE OR TWO THINGS THAT YOU ESPECIALLY LIKE ABOUT THIS TEACHER.

12. PLEASE GIVE ONE OR TWO SUGGESTIONS FOR THE IMPROVEMENT OF THIS TEACHER.

13. PLEASE NAME ONE OR TWO THINGS THAT YOU ESPECIALLY LIKE ABOUT THIS COURSE.

14. PLEASE GIVE ONE OR TWO SUGGESTIONS FOR THE IMPROVEMENT OF THIS COURSE.

Prepared by the Student Reaction Center, Division of Field Services, Western Michigan University, Kalamazoo, Michigan.

MEAN TOTAL CHANGE SCORES BY FEEDBACK CONDITION AND THEIR COMPARISON BY DUNCAN MULTIPLE RANGE TEST

Students Only	Students and Supervisors	Supervisors Only	No Feedback
-0.05	-0.39	-2.45*	-1.23*

*Significantly different from all other means, $p < .01$ (with exception of difference between second and fourth means where $p < .05$).

MEAN TOTAL CHANGE SCORES BY YEARS OF TEACHING EXPERIENCE AND SOURCES OF FEEDBACK (STUDENT VS. SUPERVISOR) AND THEIR COMPARISON BY DUNCAN MULTIPLE RANGE TEST

	Years of Experience		
	1 - 3	4 - 10	11 or more
Student Feedback	+0.04	-0.03	-0.67*
Supervisor Feedback	-1.89*	-1.11	-1.22
Mean (all 4 feedback conditions)	-1.11	-0.76	-1.17

*Significantly different from other means for that feedback condition ($p < .10$).

Figure 6. Changing Teacher Behavior as a Function of Feedback Source and Teachers' Experience Level. (From Tuckman and Oliver, 1968.)

Figure 7.

12 RULES OF EFFECTIVE FEEDBACK

(1) Feedback must involve concrete behaviors or characteristics.

(2) Feedback must provide clear, incontrovertible evidence of exactly how you appear to behave.

(3) Feedback source must be reputable and believable and intentions accepted.

(4) Feedback must be in terms you can understand and relate to.

(5) You, the feedback recipient must have a clear ideal model of what your behaviors or characteristics should be.

(6) You, the feedback recipient must also know what others' expectations of you are.

(7) You must make a commitment as to the way you would like to be.

(8) You must also make a public commitment to change.

(9) Feedback must create tension - it must be dissonant with your self-perceptions or ideals and it must be internalized.

(10) Reception of feedback must not involve more than low risk (i.e., support should be provided).

(11) Models for change and support for change must be provided.

(12) Accountability to your group must be maintained through continuing feedback.

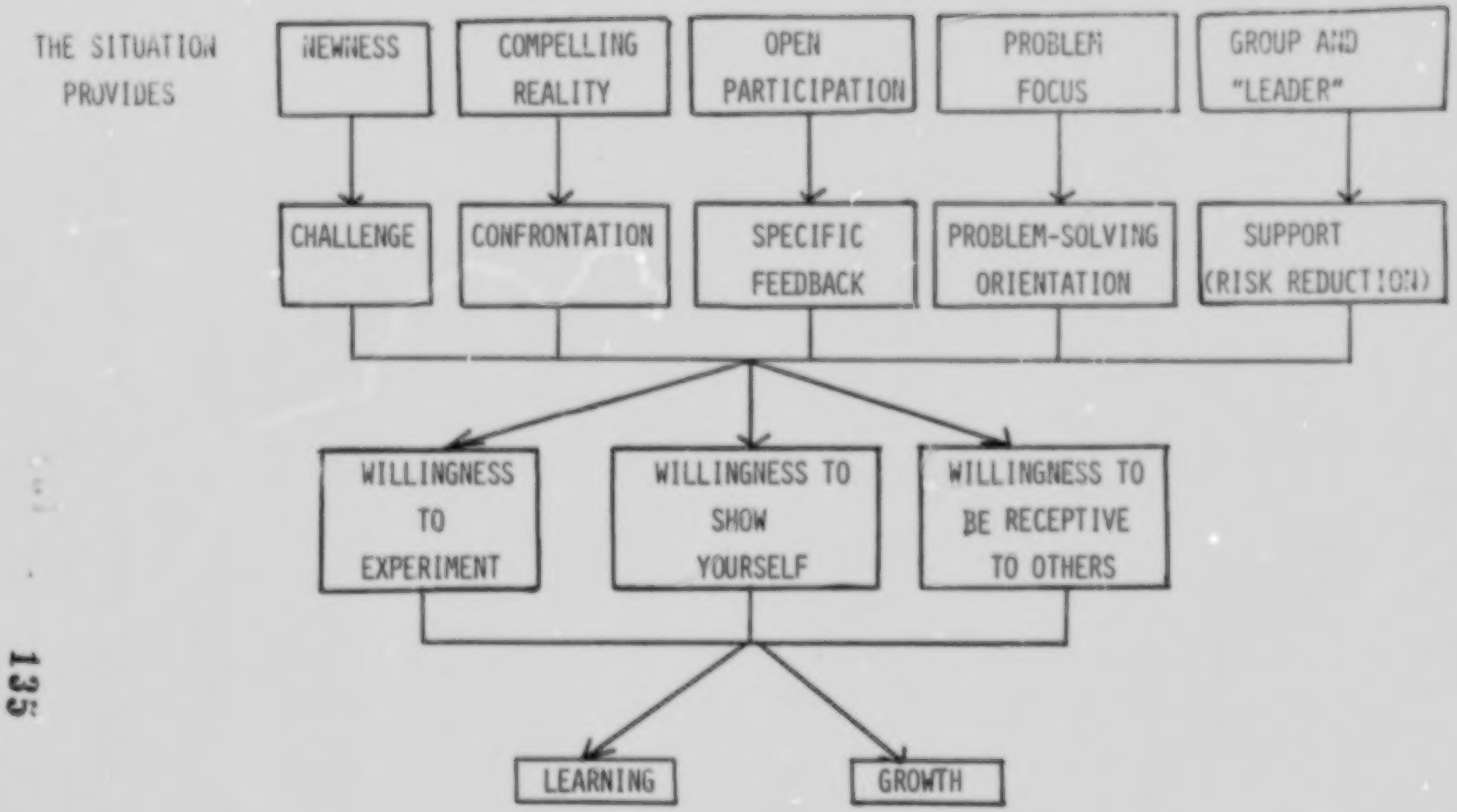


Figure 8.

THE CHANGE ENVIRONMENT

FACTOR 1

CREATIVITY

Creative-Routinized	(.84)
Imaginative-Exacting	(.83)
Experimenting-Cautious	(.81)
Original-Conventional	(.77)
Iconoclastic-Ritualistic	(.72)
Uninhibited-Inhibited	(.66)
Adventurous-Timid	(.66)
Flexible-Dogmatic	(.59)
Initiating-Different	(.52)

FACTOR 2

DYNAMISM

Outgoing Withdrawn	(.82)
Outspoken-Reserved	(.81)
Bubbly-Outlet	(.78)
Extroverted-Introverted	(.78)
Aggressive-Passive	(.76)
Assertive-Soft-Spoken	(.73)
Dominant-Submissive	(.71)
Direct-Subtle	(.65)
Buoyant-Lethargic	(.62)

FACTOR 3

ORGANIZED DEMEANOR

Systematic-Erratic	(.83)
Organized-Disorganized	(.76)
Purposeful-Capricious	(.74)
Conscientious-Flighty	(.71)
In Control-On The Run	(.62)
Observant-Preoccupied	(.58)
Resourceful-Uncertain	(.55)
Sophisticated-Naive	(.54)
Knowledgeable-Shallow	(.54)

FACTOR 4

WARMTH AND ACCEPTANCE

Warm-Cold	(.79)
Sociable-Unfriendly	(.77)
Amiable-Hostile	(.76)
Patient-Impatient	(.74)
Fair-Unfair	(.79)
Gentle-Harsh	(.69)
Accepting (People)-Critical	(.67)
Thoughtful-Inconsiderate	(.65)
Polite-Impertinent	(.64)

Figure 9. Results of the Factor Analysis of the TTFF. (Numbers in parenthese represent factor loadings; N = 84 teacher trainees)

Figure 10.

7 STEP TEACHER FEEDBACK SCHEDULE

(1) Teachers fill out ideal TTFF

(2) Teachers observe one another and fill out TTFF*

(3) Each teacher receives consensus summary statement

(4) Teachers meet as group to discuss feedback

(5) Teachers engage in strength training

(6) Teachers apply "Strengths" in regular classes

(7) Teachers observe one another again and share feedback

---

* Student judges may be used in place of teacher judges in this step.

Appendix

Teacher Observer ____________________

Observed ____________________________ Date____________________

TUCKMAN TEACHER FEEDBACK FORM

FORM A

On the following pages you will find 50 rating scales similar to the one shown below.

TALL ____: ____: ____: ____: ____: ____: ____: SHORT

You are to use all 50 scales to rate the teacher that you are observing. If you feel that the adjective tall very accurately describes the teacher, place an X in the space next to tall, as shown below.

TALL _X_: ____: ____: ____: ____: ____: ____: SHORT

If you feel that the adjective tall is somewhat descriptive of the teacher you are abserving, place an X in the second space; if slightly descriptive, place an X in the third space.

If you feel that the adjective short very accurately describes the teacher you are observing, place an X in the space next to short, as shown below.

TALL ____: ____: ____: ____: ____: ____: _X_: SHORT

If you feel that the adjective short is somewhat descriptive, place an X in the second to last space; if slightly descriptive, place an X in the third space from the right.

If you feel that either adjective is equally appropriate (or non-appropriate), place an X in the center space.

Do not place X's anywhere but in one of the seven spaces provided. Make only one X on each scale. Do not leave any blank, do not mark any more than once.

This scale will help a teacher become aware of how others see him (her). This form of feedback is essential for self-improvement. Try to be both objective and candid.

1. ORIGINAL ___: ___: ___: ___: ___: ___: ___: CONVENTIONAL
2. PASSIONATE ___: ___: ___: ___: ___: ___: ___: CONTROLLED
3. IMPERTINENT ___: ___: ___: ___: ___: ___: ___: POLITE
4. PATIENT ___: ___: ___: ___: ___: ___: ___: IMPATIENT
5. COLD ___: ___: ___: ___: ___: ___: ___: WARM
6. INITIATING ___: ___: ___: ___: ___: ___: ___: DEFERRENT
7. HOSTILE ___: ___: ___: ___: ___: ___: ___: AMIABLE
8. LIKEABLE ___: ___: ___: ___: ___: ___: ___: ALOOF
9. CREATIVE ___: ___: ___: ___: ___: ___: ___: ROUTINIZED
10. INHIBITED ___: ___: ___: ___: ___: ___: ___: UNINHIBITED
11. INCONOCLASTIC ___: ___: ___: ___: ___: ___: ___: RITUALISTIC
12. GENTLE ___: ___: ___: ___: ___: ___: ___: HARSH
13. UNFAIR ___: ___: ___: ___: ___: ___: ___: FAIR
14. BOUYANT ___: ___: ___: ___: ___: ___: ___: LETHARGIC
15. SHALLOW ___: ___: ___: ___: ___: ___: ___: KNOWLEDGEABLE
16. CAPRICIOUS ___: ___: ___: ___: ___: ___: ___: PURPOSEFUL
17. ENERGETIC ___: ___: ___: ___: ___: ___: ___: LIFELESS
18. CAUTIOUS ___: ___: ___: ___: ___: ___: ___: EXPERIMENTING
19. DISORGANIZED ___: ___: ___: ___: ___: ___: ___: ORGANIZED
20. THOUGHTFUL ___: ___: ___: ___: ___: ___: ___: INCONSIDERATE
21. UNFRIENDLY ___: ___: ___: ___: ___: ___: ___: SOCIABLE
22. RESOURCEFUL ___: ___: ___: ___: ___: ___: ___: UNCERTAIN
23. RESERVED ___: ___: ___: ___: ___: ___: ___: OUTSPOKEN
24. IMAGINATIVE ___: ___: ___: ___: ___: ___: ___: EXACTING
25. SUBTLE ___: ___: ___: ___: ___: ___: ___: DIRECT

26. ERRATIC ____: ____: ____: ____: ____: ____: ____: SYSTEMATIC

27. AGGRESSIVE ____: ____: ____: ____: ____: ____: ____: PASSIVE

28. CONCEITED ____: ____: ____: ____: ____: ____: ____: HUMBLE

29. ACCEPTING (people) ____: ____: ____: ____: ____: ____: ____: CRITICAL

30. DETACHED ____: ____: ____: ____: ____: ____: ____: EMPATHIC

31. QUIET ____: ____: ____: ____: ____: ____: ____: BUBBLY

32. AUTOCRATIC ____: ____: ____: ____: ____: ____: ____: DEMOCRATIC

33. CONTEMPLATIVE ____: ____: ____: ____: ____: ____: ____: IMPULSIVE

34. OUTGOING ____: ____: ____: ____: ____: ____: ____: WITHDRAWN

35. STUBBORN ____: ____: ____: ____: ____: ____: ____: ACCOMMODATING

36. IN CONTROL ____: ____: ____: ____: ____: ____: ____: ON THE RUN

37. FLIGHTY ____: ____: ____: ____: ____: ____: ____: CONSCIENTIOUS

38. DOMINANT ____: ____: ____: ____: ____: ____: ____: SUBMISSIVE

39. MOODY ____: ____: ____: ____: ____: ____: ____: CHEERFUL

40. OBSERVANT ____: ____: ____: ____: ____: ____: ____: PREOCCUPIED

41. EAGER ____: ____: ____: ____: ____: ____: ____: DISDAINFUL

42. INTROVERTED ____: ____: ____: ____: ____: ____: ____: EXTROVERTED

43. RELAXED ____: ____: ____: ____: ____: ____: ____: NERVOUS

44. DOGMATIC ____: ____: ____: ____: ____: ____: ____: FLEXIBLE

45. ASSERTIVE ____: ____: ____: ____: ____: ____: ____: SOFT-SPOKEN

46. EASY GOING ____: ____: ____: ____: ____: ____: ____: DEMANDING

47. TIMID ____: ____: ____: ____: ____: ____: ____: ADVENTUROUS

48. ANGRY ____: ____: ____: ____: ____: ____: ____: HAPPY

49. DOMINEERING ____: ____: ____: ____: ____: ____: ____: PERMISSIVE

50. INDIFFERENT ____: ____: ____: ____: ____: ____: ____: RESPONSIVE

Check to make sure that you have not left any scale blank, nor have marked more than one X on each scale.

Teacher Observed ______________________

Observer ______________
Date ______________

TUCKMAN TEACHER FEEDBACK FORM
FEEDBACK SUMMARY SHEET

Item Scoring

ORIGINAL 7 : 6 : 5 : 4 : 3 : 2 : 1 :CONVENTIONAL

COLD 7 : 6 : 5 : 4 : 3 : 2 : 1 :WARM

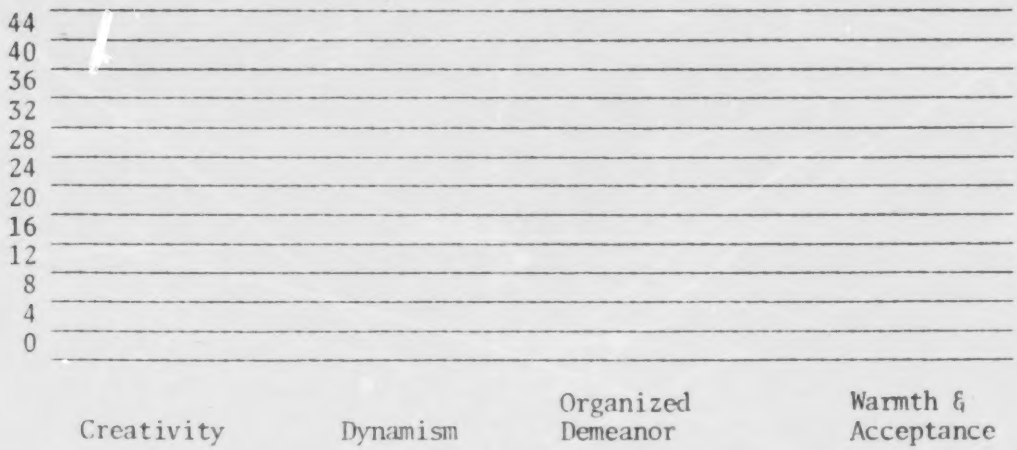
I. Creativity
item item item item   item item item
(1 + 9 + 11 + 24) - (10 + 18 + 47) + 18
(____+____+____+____) - (____+____+____) + 18 = ________

II. Dynamism (dominance & energy)
item item item item   item item item
(27 + 34 + 38 + 45) - (23 + 31 + 42) + 18
(____+____+____+____) - (____+____+____) + 18 = ________

III. Organized Demeanor (organization & control)
item item item   item item item item
(22 + 36 + 40) - (16 + 19 + 26 + 37) + 26
(____+____+____) - (____+____+____+____) + 26 = ________

IV. Warmth and Acceptance
item item item   item item item item
(4 + 12 + 29) - (5 + 7 + 13 + 21) + 26
(____+____+____) - (____+____+____+____) + 26 = ________

Profile



Bruce W. Tuckman
Copyright 1971

INSTRUMENTS FOR STUDENT EVALUATION OF FACULTY: IDEAL AND ACTUAL

Alan L. Sockloff

Temple University

The subject of my talk is instruments for student evaluation of faculty effectiveness. The reason I find it necessary to remind you of the subject is that I expect you to wonder at times whether you are listening to another talk from another conference. It is my belief that the construction of good faculty evaluation instruments involves quite a bit more than the gathering of a set of items from another unknown, unproven instrument, putting these items together in a single evening, and obtaining a dean's approval. For the construction of any instrument, there should exist a substantive philosophy and a scientific methodology as a basis. I will discuss some of the ingredients necessary for conceptualizing education, as well as some of the methodological problems that must be combatted in the construction of faculty evaluation instruments.

What are we trying to measure?

The primary difficulty in the construction of a faculty evaluation instrument stems from the complexity of such an endeavor. Besides the many sources of evaluation and the many purposes that can be served by the evaluations, the fact that faculty responsibility covers a broad domain suggests that there are also many facets of faculty activity that can be evaluated. In a large industry in which the primary goal is the dollar, the assembly line worker can be evaluated by a foreman via the application of a single numeric measure of his productivity, i.e., the number of windshield wipers attached daily. Can the same be done for a faculty member in an institution of higher education?

The three major sources of faculty evaluation are administrators, faculty colleagues, and students, and the three major purposes of evaluation are for making administrative decisions, faculty feedback, and student use. Interestingly enough, a parallel exists between the sources and the purposes. Does this imply that we can take advantage of the parallel and have administrators evaluate faculty for administrative decisions, colleagues evaluate faculty for feedback, and students evaluate faculty for student use? The limitations of this approach should be obvious. The limited perspectives of the three source groups defeat the purposes of the evaluations. It is doubtful that in a "natural" environment each of the source groups would have the same opportunities to observe the same characteristics and behaviors. This is particularly true because the responsibilities of faculty members are quite complex.

It can be safely stated that faculty responsibilities are equivalent to, and encompass, the goals of higher education. At a very general level, the goal of higher education is education. Without trying to get involved in one of those ad nauseum discussions on the meaning of education, I would like to present an over-generalized definition that should arouse little disagreement. "Education" is defined here as a "process of change in some desired direction, where this direction involves both short-term and long-term objectives." Although recognizing education as a process, I would like to also treat it as a goal. Education necessarily involves two components, a learner and a stimulator, where the stimulator is a stimulus set external to the learner, e.g., a teacher. The interest here is not in the separate components, but rather in the interaction between the components.

Fortunately, the subject of this conference pertains to evaluation of faculty effectiveness by students. Thus, concentrating solely upon student evaluation delimits our problem somewhat, but not drastically so. In terms of what students can observe and evaluate, a faculty member's responsibilities include course instruction, modelling of an adult and citizen, counseling, scholarship, and so on. I must admit that this list covers broad areas, and is not very comprehensive. Nevertheless, I am contending that those aspects of faculty effectiveness that can be evaluated by students include more than the ability to babble nonsense (which could be read in a textbook) for 10 to 20 hours per week to a group of attentive listeners.

An extension of some old notions.

One of the common notions thrown around these days is that faculty effectiveness can be measured on the basis of the so-called measures of "learning," -- the achievement tests. Let's take a look at one of the bases of this notion, an elementary school model, and try to determine whether the generalization of such an approach to faculty evaluation in higher education is feasible.

In the old, traditional sense of an elementary school education, in which the desired goal is a grasp of the rudiments, the 3 R's, evaluation of teacher effectiveness is quite straightforward. To a great extent, the short-term objectives in this educational model include the learning of the 3 R's as a basis for later learning. Assuming random assignment of students to teachers, the fact that each child is subjected mainly to one teacher for a full school year allows us to evaluate and compare teachers according to changes in the scores of their students on annual, common achievement examinations.

Now, would such an evaluation model fit within the context of higher education? To answer this question, I propose here to construct the HEI, the Higher Education Instrument, through the combined efforts of the world's great and not-so-great substantive experts and psychometricians. The HEI will consist of many subtests, one for every short-term objective (i.e., every conceivable course ever taken in higher education) and one for every long-term objective (e.g., problem-solving, intellectual independence, enlightened citizenship, emotional maturity, etc.). The HEI could be administered during the summer prior to entering a degree program and directly after completing the other program requirements. I suspect that we would also have to set aside weekends during these summers in order to motivate the testees through the administration of pep talks and weekly supplies of pep pills. From the HEI results, every teacher could be evaluated on the basis of pre-post difference scores on the subtests of the students that he taught.

The flaws in such an undertaking should be obvious--the HEI is an absurd caricature. Proponents of such an approach, on a somewhat less grandiose scale, would have us believe that this is the only valid approach. They would argue that a scaled-down HEI would allow us to attribute "learning" of the students to the teachers. But, if the purpose of such testing is the evaluation of faculty effectiveness, then I would argue that there are more efficient methods to achieve this purpose.

An hypothetical model of education.

Let's assume a hypothetical multi-dimensional space. The axes of this space have labels corresponding to the objectives of education such as Knowledge, Understanding, Problem-Solving, Intellectual, Independence, Emotional Maturity, etc. Somewhere in this space, we have two points, one designated Learner and the other designated Education, where both points can be located by distances from the origin along the various axes. The exact location of the point Education is _not a matter of fact_, but a matter of decision on the part of the institution that defines education in terms of some proportional balance of objectives.

I am tentatively positing a generalized model of Education. According to this model, I am contending that an effective faculty member is one who, in comparison to some standard, by appropriate stimulation propels Learners closer to the point Education. Learning is the distance moved by a Learner directly toward Education. It appears, then, that all we must do to evaluate a faculty member is to measure the Learning of students stimulated by him. Herein lies our problem. The model and the points are hypothetical, the objectives are hypothetical constructs, and distances propelled along the axes corresponding to the various objectives are not directly measurable. Since we recognize that we cannot directly measure either the distances propelled along the axes or the distance moved directly toward the point Education, then perhaps we can measure other quantitites that are estimates of (and correlated with) the distances propelled. But, how can we determine that we are accomplishing this in our measures? Or, rather, how can we validate our measures?

Construct validation and measures of faculty effectiveness.

Construct validation, as espoused by Cronbach and Meehl (1955), arose as a method for validating measures in situations in which the classical approach to criterion-oriented validation is inappropriate. The logic of criterion-oriented validation generally involves the computation of a correlation coefficient between the scores of a given test, be it personality, attitudinal, interest, achievement, or whatever, with scores on some criterion measure. The distinction between the measure derived from the given test and the criterion is simply a matter of cost: money, time, subject cooperation, etc. The criterion is more costly to measure directly, and it is more expedient to have some estimate of the criterion that can be more easily measured. However, when the criterion itself cannot be measured directly, and no single correlation coefficient can be calculated as an estimate of the validity of some given measure, the classical theory of validity becomes inadequate.

According to the logic of construct validation methodology, there are hypothetical constructs that are not directly measurable and constructs that are directly measurable. In addition, there exists a nomological net that consists of a set of "laws" interrelating all of the constructs. After the nomological net, or model, has been hypothesized, research is used to assess the relationships specified by the model, as well as to suggest changes in the model on the basis of empirical evidence.

An approach that can be used to represent the construct validity method was proposed by Campbell and Fiske (1959), the multitrait-multimethod matrix. This approach makes use of two types of validity--convergent and discriminant. Whereas convergent validity requires that a measure of a particular construct be highly correlated with other, independently obtained measures of that construct, discriminant validity requires that the measure of that particular construct have lower correlations with measures of other constructs.

Our interests concern the distance moved by the Learner toward the point Education (and the distances moved along the axes toward the various objectives), as stimulated by the Teacher. Clearly, we are dealing here with hypothetical constructs for which we would like to have accurate measures. Let's imagine that we constructed a measure of the hypothetical Learning distance, and we call this measure Faculty Effectiveness. In terms of the Campbell and Fiske approach, our measure Faculty Effectiveness has convergent validity if it is highly correlated with other measures of the hypothetical Learning distance, and our measure Faculty Effectiveness has discriminant validity if its correlations with measures of other hypothetical constructs are quite a bit lower.

A simplified example should help to clarify both aspects of construct validity through the multitrait-multimethod matrix. I could have theatrical directors sit in with students in several classes for a semester, and have both groups of observers rate teachers in terms of Faculty Effectiveness and Acting Potential. I could then calculate group means and 6 correlations between measures across group means of the different classes. The interest of this little study is to help validate the Faculty Effectiveness measure, and the results I would not mind obtaining are the following. I would like my highest correlations to be between students' and directors' Faculty Effectiveness measures (and between students' and directors' Acting Potential measures). I would also like to find the other correlations substantially lower. If, in fact, I found that my highest correlations were between students' Faculty Effectiveness and Acting Potential measures, I might have to conclude that unless I could find some way of conceptualizing acting as a measure of the hypothetical construct Learning, my Faculty Effectiveness measure is doomed and back to the drawing board I would go.

The point that I want to make here is that tools exist for the validation of instruments and their items. Admittedly, such tools lead to the establishment of long-term research programs, but until many of the constructs, both hypothetical and measurable, can be specified in terms of their interrelationships, there cannot be a satisfactory instrument for student evaluation of faculty effectiveness. Simply stated, the "ideal" instrument consists of *measures that are valid with respect to the construct "Learning as stimulated* by the Teacher."

Proponents of the HEI approach insist that when measures of student evaluation of teaching either fail to show significant relationships with achievement tests scores or show significant relationships in the opposite direction (i.e., students assigning high ratings to the teacher are those who received low scores on achievement tests), that this invalidates the use of student ratings. But, when neither measure has been validated against the hypothetical Learning construct, such results really show nothing. All too often, the means are confused with the end-products, and associated with this erroneous reasoning is the belief that an achievement test score is itself the hypothetical construct Learning. A positive feature of the achievement test approach is that it may well lead to reasonable estimates of a Learning construct, without suffering too severely from response biases that are so typical of rating instruments. But, surely, the objectives defining Education are not likely to all be Knowledge-related. Furthermore, the standardization of evaluation procedures brought about by student ratings is a desirable feature. A single rating instrument, with items of proven validity, can be more conveniently administered than achievement tests and would allow comparisons across courses, departments, or colleges.

The question of the student's ability to evaluate faculty is often raised: Who are students to judge? This is a fair question because on one hand we are asking the student to go through the process of education, and on the other hand we are asking the student to objectively judge either his own educational progress or the characteristics of his teachers that lead to his educational progress. The answer is simple. If the characteristics of a good teacher can be defined and validated with respect to the construct Learning, and are observable and accurately rateable, then student evaluation can be a practicable solution for the various purposes of evaluating faculty effectiveness.

Some principles in constructing rating scales.

The preceding discussion of construct validity and its relationship to student evaluation of faculty effectiveness presupposed both the construction of the measures and satisfactorily high reliability of the measures. I will briefly review the logic and considerations involved in the construction of rating items, and this will be followed by a rather brief note on reliability. I am avoiding any mention of open-ended questions, since the major use I see for this type of question is for faculty feedback and self-diagnosis.

The most common technique used in faculty evaluation instruments involves rating scales. Although not necessarily in terms of item format, but in terms of purpose, an important distinction exists between rating scales and attitude scales. The purpose behind a rating scale is to objectively describe some external object, whereas the purpose behind an attitude scale is to subjectively describe one's reactions to, or attitude toward, that external object. With regard to faculty evaluation instruments, this distinction is sometimes clouded. I do think that we should be more interested in rating the teacher than in measuring students' attitudes toward that teacher. The reasoning behind this is that objective ratings of behaviors and characteristics of the teacher should have a smaller, more "controllable" set of biases than students' subjective attitudes toward that teacher.

For the most part, two types of rating items have been used in faculty evaluation instruments: numeric ratings and graphic ratings. Both item types involve a stem, which is a statement regarding a characteristic of the teacher, and a series of cues, which are ordered adjective and/or adverb phrases or words. For the numeric rating item, numbers are frequently associated with the cues, and the respondent is asked to mark one response on the questionnaire or to record the number associated with the cue on a scorable answer sheet or punch card. For the graphic rating item, there exists a response continuum, which is usually a segmented or continuous line (more likely horizontal), with cues to identify regions along that line. The respondent is asked to mark some point on that line. Although graphic rating items are more easily administered, numeric rating items are more easily scored.

According to Guilford (1954), the following are some guidelines for the construction of stems. The stems should describe traits, qualities, or behaviors that:

(1) are objective and specific,

(2) are not a composite of independent traits, qualities, or behaviors

(3) refer to a single type of activity or its results,

(4) are judged on the basis of present or past performance, not on future promise.

In addition,

(5) stems should not contain cues.

Furthermore, according to Guilford (1954), the following are guidelines for the construction of cues. Cues should:

(1) be short and unambiguous,

(2) be consistent with the stem and other cues for that stem,

(3) have a precise, short range,

(4) have varied language with respect to a single stem,

(5) avoid ethical, moral, or social evaluations,

(6) not be similar across stems (i.e., non-common sets of cues).

In constructing responses for a numeric rating item, there are additional considerations regarding the number of responses in the scale and characteristics of the cues. The number of responses should be such that the respondent can discriminate the gradations between responses with the aid of the cues. For statistical reasons (minimizing platykurtosis and skew), it is preferable to have the cue anchors (the two most extreme cues) sufficiently extreme that they will draw few responses; thus, if there are k possible responses, there would be k-2 functional (most used) responses. The 5-point scale item is fairly popular, and if the anchor responses were designed to be used rarely, this would leave only 3 functional responses. In this case, the amount of lost information from a functional 3-response scale depends on the extent to which the respondent could have made finer discriminations. In many of the faculty evaluation instruments that I have seen, there is a built-in functional asymmetry insofar as the negative anchor cue is quite extreme and has little drawing power, while the positive anchor cue is not so extreme and has a stronger drawing power, thus skewing the item response distributions.

It may be desirable for the responses to be subjectively equidistant, but this should not be done at the expense of truncating the range of functional responses. If reasonably equal subjective response intervals are desirable, it may be necessary to cue all of the responses, not just the anchor responses. Another good reason for trying to cue all of the responses is that the lack of cues may arouse ambiguity, which can lead to the operation of response biases in the functional range of the scale. And, last, for statistical reasons, the choice of cues should be dictated by efforts to have the mean response across instructors centered in the middle of the scale.

The real bugaboo in rating scales is the operation of response biases or errors. If care isn't taken in writing items and training raters, responses may contain little more than the effects of bias. According to Guilford (1954), there are at least 6 general categories of response bias: logical, proximity, central tendency, leniency, contrast, and halo. Although all of these response biases can be attributed to personal idiosyncrasies of the raters, the first three categories can be considered non-interpersonal. Logical bias is the tendency to give similar ratings on items that look similar. Proximity bias is the tendency to give similar ratings on neighboring items in the instrument. Central tendency bias is the tendency to give central ratings rather than extreme ratings.

The remaining three categories of response bias may operate when other people are being rated. Leniency bias is simply a characteristic of people-as-raters--some people are just more lenient than others. Contrast bias is the tendency to rate other people as being opposite from oneself. Halo bias, perhaps the most serious of biases in inter-personal ratings, represents a generalization of an overall subjective feeling toward the person being rated to the rating of specific qualities of that person.

Of the many faculty evaluation instruments and individual faculty summaries that I have seen, I think that the operation of the central tendency and contrast biases are, if not minimal, far outweighed by the effects of the leniency, logical, and halo biases. I think that most students are unwilling to be overly critical of their teachers, and this may be due in part to their suspicions about the anonymity of their responses. Further, the use of poor, relatively global-type items seems to almost demand personal response bias rather than objectivity. For this reason, I suspect that a good actor who assigns high grades and stimulates little in the way of Learning can fare pretty well on instruments consisting of items that violate most of the guidelines.

The operation of response biases are particularly problematic when it comes time to assess the reliability of an instrument. By reliability, I mean stability or consistency of measurement. The appeal of highly reliable instruments and the efforts expended toward the development of reliable instruments stems from the psychometric axiom that reliability sets a ceiling on validity. The temptation of the researcher who is both aware of the axiom and aware of the difficulties in assessing validity may be the following: "Well,...things can't be that bad if my reliabilities are so high." The problem is simply that reliability is a necessary condition, but not a sufficient condition, for validity.

Since the interpersonal response biases (leniency, contrast, and halo) can be thought of as relatively enduring traits of the raters with respect to the rating of a particular teacher, the variance attributable to these response biases is included with the variance attributable to true scores, thus exaggerating reliability estimates. Given a set of poor, ambiguous, global items, with absolutely no validity with respect to Learning, I am certain that I could provide you with reliability coefficients in the .80's or even the .90's. Until it can be demonstrated that response bias has been minimized or statistically controlled, we are wasting our time calculating reliability coefficients.

Some issues in item and instrument construction.

A great deal of latitude exists in the methods for constructing faculty evaluation instruments. Considering this latitude, it is not very surprising that different researchers achieve different, and sometimes contrary, results in research relating student ratings of faculty effectiveness to other measures. Until such time that the "ideal" instrument is developed, some of the research differences will just have to be tolerated and tentatively attributed to instrument differences or sample differences.

There are, however, a few issues relating to instrument construction that are, if not resolvable, at least worth discussing at this time. The tack that I shall take in the discussion of these issues is based on my conception of the principles of common sense in conjunction with the psychometric properties of good items. Discussion of these issues is critical if the instruments we construct are to ever withstand the rigors of validity testing. The following issues will be discussed briefly: the selection of potentially valid items; behavioral vs. global items; the use of composite scales; the use of common cues; the choice of response continuum; the use of normative data; the rated object; and traditional vs. progressive items.

The selection of potentially valid items is ideally done through a model of learning in higher education. Since there aren't too many models being kicked around these days, some other selection methods are needed. Critical incident techniques and open-ended requests for traits seem to be fairly successful methods for gathering items. The most popular method of selecting items is the "prestige library" method, the borrowing of items from popular, prestigious instruments. A very necessary, but often overlooked, step for items that have been selected through means other than a model is that of "ORA" consensual validation: observability by the target source group; rateability by the target source group; and acceptability by other source groups as measures that are potentially related to a Learning construct. The ORA consensual validation should give a comfortable headstart on eventual construct validation.

The behavioral vs. global item issue concerns the complexity of the behaviors rated. The following two stems are typical of the two extremes: "The teacher made use of illustrations to get across difficult points"; and

"Overall, the teacher was..." For several reasons, mostly having to do with the properties of global items, I believe the behavioral items to be superior. First, global items are non-specific and somewhat ambiguous, thus leading to the operation of response bias, particularly halo bias. Second, in general, there is a little reason to have great faith in the reliability, and validity, of single items. Decision-makers (students deciding which courses to take and administrators deciding who to promote), when viewing the faculty evaluation results of individual teachers, tend to search out one or two "comprehensive" items as a basis for their decision. The fact that these one or two "comprehensive" items are global items, and are overly subject to response bias, suggests that decisions regarding a professor's performance are more likely based on the extent to which he is liked, not necessarily the extent to which he is a good teacher. Third, global items have a little diagnostic value, and fourth, behavioral items fail within the realm of objectively observable.

Besides the diagnostic value and the better capabilities of minimizing response bias, an additional positive feature of behavioral items involves their potential use in composite scales. On the basis of factor analytic, clustering, or even rational, techniques, various groupings of items can be summated to create composite scales. Assuming that the behavioral items are good items, the inherent advantages of scale scores include high reliability (and potential validity), as well as comprehensiveness.

The use of common cues with common scale directions has some interesting ramifications. Considering that students may evaluate several faculty in any given semester, a long set of items with cues unique to each item may lead to boredom, fatigue, and eventually large doses of response bias. One alternative is to use a short set of behavioral items with unique cues, but the set of items itself could not be very comprehensive. Another alternative is to use a short set of global items with unique cues. And still, another alternative is to use a large set of behavioral items, making use of common cues with common scale directions to facilitate the administration of the instrument. But, this might not be a good idea either, since the common cues may lead to the non-use of cues by the raters and the operation of response bias. Since none of the alternatives are satisfactory, we will have to await methodological research considering these questions.

Those who dare fate by using common cues sometimes do so for reasons of expediency, e.g., the restricted area on optical scan sheets sometimes forces instrument constructors to use common cues if the stems, cues, and response areas are to be on a single sheet. If common cues are to be used, what are the appropriate continua underlying these cues? The following examples of underlying cue continua also include my perceptions of their limitations in terms of introducing response bias. The "agreement-disagreement" continuum suggests the subjectivity of attitudes rather than the objectivity of rating. Other continua, such as a "success" continuum, are highly value-laden and may well lead to the same result. At first blush, the "frequency" continuum appears to have some nice objective properties, but since it may be difficult to fit every stem to ratings in terms of ranges of frequency of occurrence, this too may lead to response bias, particularly for stems that do not comfortably fit the cues. Various other continua, such as a "characteristic-uncharacteristic" continuum, may turn out to be ambiguous and, thus, ignored for cuing responses. I think that what we have here is another open area for methodological research.

With regard to use of normative data, I fail to see any issue--normative data is an absolute necessity. Since the numeric values of item ratings are completely arbitrary, with no zero point, and since the assignment of numeric ratings is very much influenced by the particular set of cues for each item, the only meaning that can be attached to item mean scores or frequencies or to composite scale mean scores derives from comparison with some kind of standard, such as a normative group. In addition, flexibility can be gained by using different types of normative data, such as college norms, department norms, student class norms, or even individual faculty norms.

Student evaluation of faculty effectiveness has been made through the rating of three objects: the teacher, the course, or the students' own educational development. Ideally, effective faculty offer good courses in which students learn. But, if the ratings of the three objects do not jibe, what does this mean? Hartley and Hogan (1972) factor analyzed teacher-course description ratings adapted from McKeachie's form along with the student's ratings of their own self-development. Hartley and Hogan's results revealed factors that were defined by either teacher-course descriptions or by self-development items, but generally not by both types of items. These results raise an interesting issue. Although the self-development approach would seem to be a good method for ridding response biases with respect to the teacher (and course), it may provide little more than a vehicle for the operation of a completely different set of response biases, self-perception response biases. Unquestionably, this is another one of those issues that is in need of clarification from research.

The last issue I would like to tackle is that of traditional vs. progressive items. Items obtained by the "prestige library" method tend to be traditional items. By traditional items, I mean items that are generally appropriate for most varieties of teaching situations. Progressive items are those that allow for differing, non-standard types of behaviors and approaches that may be effective as teaching and learning techniques. Since not all teachers experiment with teaching devices and techniques, with regard to instrument construction, there should be some concern that progressive teachers would not receive low ratings on items that represent traditional behaviors which were replaced by that teacher. For example, the teacher who found that students in his courses gave more creative responses on examinations if he did not tell them how to study for the course would not fare well on the following item used with "frequency" cues: "The teacher gave advice on how to study for the course." A not-so-pleasant alternative approach to avoid this problem would be the use of global items.

A jaundiced view of what people do

Dick Riley and I were curious about how people actually constructed and used instruments for faculty evaluation by students. Although some of the requests were lost in the mail, we wrote just under 3,000 American institutions of higher education, requesting information about uses, financial support, sources of items, and methods, as well as copies of instruments, exemplary individual summary sheets, and technical reports describing the construction of the instruments. Our questionnaires were sent to the highest ranking academic administrator whom we thought would be concerned with student evaluation of faculty. We have received around 900 responses--this higher-than-expected return rate may have resulted from our promise to send copies of our report to the returnees.

Our responses came from a variety of institutions with respect to type (university, 4-year college, 2-year college, technical schools, post-graduate), size, sex (single sex, coeducational), and control (private, public). Around 500 instruments were sent. In addition to a somewhat smaller number of exemplary individual teacher summaries, the number of technical reports describing construction and research on the instruments was quite small. Being optimists, we just assumed that someone forgot to send along the reports. On the other hand, we'll never know whether the technical reports that we failed to receive actually exist, unless we were willing to go through some extensive follow-up procedures.

The large majority of the instruments that we received were used primarily for faculty feedback purposes, and to a lesser extent, for administrative decisions and student perusal. It was, however, mildly disturbing to learn that in more than one-half of the cases, the individual faculty summaries were seen by decision makers (administrators, department chairmen, and students).

The modal, typical instrument contained between 11 and 30 items, largely derived from other instruments--some by the "prestige library" method and others by the "not-so-prestige library" method. The instruments typically contained professor items, course items, global items, and open-ended questions. Student development items were used, but did not seem to have the popularity of professor items. Most of the instruments were mimeographed, with responses to be marked on the instrument itself. Norms were used in conjunction with around 10% of the instruments.

With a few exceptions, my own undocumented, global rating of the instruments would be the negative anchor on a 5-point scale. Item stems contained statements about many unobservable characteristics of faculty and courses or characteristics that should not be evaluated by college students. In addition, the combination of emotionally loaded stems and cues that are suggestive of attitudinal or evaluative judgments seemed to ask for responses that contain little more than bias. As far as I could see, rare is the instrument that hasn't violated at least one of the principles of good item writing. Perhaps, if the instruments had been researched early in their development, a good 99% would not be around today.

Much of the blame for these conditions should be placed on the colleges and universities themselves. Although acknowledging the need for student input in the decision-making, these institutions have certainly tolerated, but not encouraged, student evaluation of faculty. The reasoning seems to be as follows: If everyone can agree on the inferiority of the bulk of the available instruments, then no one really has to take them seriously.

In conclusion, I have tried to show that the methodology and technology are available for the construction of instruments. Even though, by definition, the "ideal" instrument may never be constructed, the process of striving for this goal should lead to vast improvements over the status quo.

References

Campbell, D. T., and Fiske, D. W. Convergent and discriminant validation by the multitrait-multimethod matrix. Psychological Bulletin, 1959, 56, 81-105.

Cronbach, L. J., and Meehl, P. E. Construct validity in psychological tests. Psychological Bulletin, 1955, 52, 281-302.

Guilford, J. P. Psychometric methods. New York: McGraw-Hill, 1954

Hartley, E. L., and Hogan, T. P. Some additional factors in student evaluation of courses. American Educational Research Journal, 1972, 9, 241-250.

THE KANSAS STATE UNIVERSITY PROGRAM
FOR ASSESSING AND IMPROVING INSTRUCTIONAL EFFECTIVENESS

Donald P. Hoyt

Kansas State Universtiy

Although my purpose is to describe the teacher evaluation program in operation at Kansas State University, a brief review of its history is necessary both to set the climate and to provide a rationale. The program had its inception in October, 1968, during my second month as Kansas State's Director of Educational Research. I was not as naive politically as that may sound; while I was fully aware that faculty evaluation would inevitably become a central concern of our Office, I intended to spend my first year or two on less controversial and threatening problems. I felt a good program required the trust of the faculty, and gaining that trust would take time.

This assessment, while absolutely correct, became less persusaive as a deterrent when, over the course of two weeks, contingents representing faculty-student committees on instruction in three different colleges sought my advice on developing their own devices for appraising instructional effectiveness. Given the alternative of having multiple amateurish efforts whose quality would be questionable and whose administrative procedures would be chaotic, I concluded that the potential dangers would be less if I made a serious (though premature) attempt at appraising teaching effectiveness. With an interpersonal touch they don't teach in graduate school, I successfully inveigled each committee into requesting that I design a system that would meet the needs of all three.

The first problem which had to be resolved concerned purpose. There was universal agreement that "improving teaching effectiveness" should be a major thrust. In addition, there was considerable sentiment among students that the program should also produce results which would help them select courses and instructors. Students also tended to be sympathetic to my bias that results should be used when making decisions related to salary, promotion, and tenure. A violent faculty reaction to both of these ideas made it clear that the only premise under which we could proceed was that the sole purpose of the program would be the improvement of teaching. I helped convince the student representatives to accept this not only because it was worthy but also because I felt that, if the program succeeded in this way, progress on the other purposes might become feasible in later years. As it turned out, this expectation was realized.

A number of faculty members and students served as consultants. The faculty were particularly helpful. I began by presenting them with lists of items describing teaching behavior, stolen from various sources. I asked them to indicate which of the items were especially descriptive of good teaching. While most of the faculty consultants were courteous and made constructive comments, two or three of the most hostile ones had the most positive effect on my thinking. One went to considerable trouble to show how each of the items he was reviewing could be symptomatic of inferior as well as superior teaching (e.g., "The teacher who lets students discuss the fact that 2 plus 2 equals 4 wastes his time and that of his students;" "Well-organized garbage still smells; and disorganized pearls are still precious;" "Lovin' 'em don't learn 'em; the price of your popularity is their ignorance.") Another astutely pointed out that any attempt to describe the ideal teacher by a standard set of items was doomed to failure because what was effective was dependent on the situation.

I felt a little foolish to have the obvious pointed out so clearly--techniques that work well with large classes don't necessarily work with small classes; and faculty members who are trying to get across solid factual content may have to use methods quite different from those who are trying to stimulate students to examine their motives or values.

These critiques led to the most important decision in designing our program and to the feature which most distinguishes it from others I have examined. I refer to the decision as to how teaching effectiveness should be defined. I could see no way to define it by describing any single role model. Rather, my most persuasive critics were saying, indirectly, that good teaching is recognized by its products. Examine what happened to the students and you'll know if the faculty member was effective or not.

When I asked my consultants to respond to that reasoning, I found no serious objections. What they did say was that there was no way to design a system based on this logic because the outcomes expected in each course would be different. Clearly, the effectiveness with which music appreciation was taught would require different measures of student outcomes than would be needed for a course in thermodynamics.

While I recognized the difficulties, earlier experiences convinced me that they may not be insurmountable. So a new tack was taken. Using the taxonomies (Bloom, 1956; Krathwohl, et. al., 1956) and some stimulating work by Deshpande and Webb (1968), I tried my hand at developing a list of general objectives which could be used to describe the purpose of any course. After several committee meetings and considerable debate, I was left with a list of eight objectives which seemed to do reasonable justice to the literature and to the suggestions of my consultants. The latter agreed that, by supplying importance ratings to these objectives, faculty members could provide a profile of their objectives which would adequately describe their courses.

Now all we had to do was measure progress on these objectives. If we had reasonably adequate measures of student progress, we could combine them with the instructor's rating of importance to obtain an evaluation of teaching effectiveness which took into account the unique pattern of objectives for each course. I recalled some earlier personal experiences in the development of empirical measuring devices which had been the source of some embarrassment. For example, after spending several thousand dollars of the Hill Family Foundation's money to measure anxiety, the most potent item our research uncovered was "I feel anxious about someone or something almost all the time:" And in developing an alcoholism scale, our best item was "I have used alcohol excessively in the past." These experiences encouraged me to try the simplest, most direct approach; namely, to ask the student how much progress he made. I had been involved with and knew of a number of studies which suggested the value of self-ratings (e.g. Holland & Lutz, 1968; Keefer, 1965; Walsh, 1967). And Nate Gage provided an inadvertent boost to my confidence when, a year earlier, he told a seminar I was teaching that student ratings of their knowledge gain correlated substantially with objectively measured gain in some of the mini-unit studies he was conducting at Stanford. I finally found a study by Soloman, Rosenberg, and Bezdek (1968) that reported findings which strongly supported my bias. This was enough intellectual armament to win approval from my consultants for a trial.

The rest of the technical history is quite routine. I devised an instrument which allowed students to rate their progress on the eight objectives. It also contained a number of items, mostly plagiarized (Isaacson, et. al., 1964; Soloman, 1966; Whitlock, 1966), to describe the instructor's behavior and the course. A separate form was used to collect instructor ratings of the importance of each objective. After intensive politicking, most faculty members "volunteered" to participate in the developmental run which was conducted in the second semester of 1968-69.

With results from well over 700 classes, there was plenty of data to divide the classes into "developmental" and "cross-validation" groups. Sixteen partially overlapping developmental subgroups were formed by sorting classes into two sizes (50 or more students and less than 30) and into one or more groups based upon instructor objectives. For example, one subgroup contained all classes enrolling fewer than 30 students where the instructor had rated the objective concerned with gaining factual knowledge as "essential". Similar subgroups were formed for large and small classes stressing each of the seven remaining objectives. Classes within a subgroup were then assigned to one of six "progress" categories on the basis of the average rating of student progress on the objective in question. Then, statistical analyses were performed to determine how descriptions of instructors whose classes made "much progress" differed from those where student progress ratings were low. Resulting scales were then cross-validated.

In the end, a few items were found to be characteristics of effective teaching regardless of the objectives being sought or the size of the class. A few other were predictive of progress ratings in small classes but were unrelated to progress in large classes; the reverse was also true. And a few items didn't differentiate among progress groups on any criterion. But for the most part, the contention of my early critics was substantiated.

The particular teaching behaviors which were related to student progress were different for each objective and for large and small classes. Rather than one model of teaching effectiveness, we had developed sixteen.

Other statistics were also encouraging. In cross-validation studies, the 16 special scales correlated from .50 to .83 with class progress ratings with average values of .68 and .62 for large and small classes, respectively. Reliability figures were generally over .90 for classes of 15 or more students. Costs averaged about $3.00 per class, and these were covered by the U. S. Office of Education which had generously funded the effort (Hoyt, 1969).

We soon discovered that conducting a good study doesn't guarantee the implementation of a good program. My request for supplemental funds to establish a service program based on our research was referred to the Faculty Senate which expressed the sentiment that the University had more pressing needs. It was finally agreed that faculty members who requested the use of our device could be accommodated if the dean of his college would pay the computing center costs. This procedure nearly resulted in the stillbirth of the program; fortunately one dean not only provided blanket authorization for his faculty but made it clear that he believed volunteering to participate was a positive thing to do. This kept the program alive, but at a very reduced level. Results for approximately 80 classes were processed in 1969-70.

By happy accident, the Council of Academic Deans had voted a year earlier to establish a new office of faculty development. When it became apparent that enrollment increases would merit a number of new positions in 1970-71, a search committee was activated. As luck would have it, a popular member of our College of Education who had served as a teaching consultant to many departments expressed an interest in the position. His appointment in the fall of 1970-71 marked the end of vigorous resistance to the program. As best I can interpret the situation, the changed atmosphere was due partly to the one year cooling-off period, partly to the idea that the teaching improvement program would include both appraisal and expert consultation, and partly to the highly positive image of the faculty member appointed.

In any event, through his office the program has been offered on a volunteer, confidential basis every semester since. The number of participants has steadily grown from about 250 classes in the fall of 1970 to over 400 last fall. While instructional improvement has remained the program's major thrust, over 90 percent of the participants last fall agreed to release selected parts of the report for publication by a Student Senate committee. And by recent action of the Faculty Senate, results from this or similar devices must be made available to the department head before reappointment decisions are made. In addition, the instrument plays a major role in selecting winners of the outstanding teacher cash awards. Three years after its traumatic birth, the program is thriving and finding broad application on our campus.

Teaching improvement is the major purpose of the program. In a recent study of changes of scores over one and two year periods, there is the suggestion that the program has enjoyed at least minimal success. Retest scores for the same course-instructor combinations showed significant gains on both student progress ratings and on a number of instructional methods scores

Let me describe more specifically how the evaluation is used to improve instruction. You will recall that the basic research effort resulted in identifying 16 lists of items describing the classroom behaviors characteristic of the most successful teachers (i.e., those whose students reported the most progress on a given objective). These lists of relevant items have become the focal point of our efforts to improve instructional procedures.

Typically, the process goes like this. A computer report and interpretative manual is sent to the participating faculty member. He reacts with confusion, diasppointment, or curiosity and accepts our invitation to attend a group interpretation meeting. There he is asked to identify the areas of greatest concern by comparing his importance ratings with student progress ratings; appropriate norms are used and most instructors attending these sessions find at least one important objective where student progress ratings were below average. When such an objective has been identified, the faculty member is asked to review the particular teacher behavior items which were positively related to gains on this objective. He is shown how most teachers are rated on these items and his printout shows how he was rated. Invariably, his rating will be unusually low on a few of these crucial items. Presumably, these items will form the basis for his self-improvement efforts.

What happens then is not highly predictable. Some seem to resolve to do better and let it go at that. Others may arrange to attend one of the special seminars on teaching procedures conducted by the Director of Educational Improvement. Still others make individual appointments with the Director and embark on individual improvement programs of various degrees of intensity. Figure 1 shows the "Before" and "After" results for one faculty member who embarked on a serious, time consuming self-improvement program under the supervision of the Director.

Although the program is alive and well, the instrument was thoroughly revised in the fall of 1972-73. The revision reflects both the criticisms of experienced users and re-thinking on our part of how our improvement purposes could be best served. Let me give you an indication of the changes we felt were required without boring you with details.

1. The list of objectives was revised and expanded. As users became more acquainted with the key role objectives play in our evaluation process, they became more articulate about their purposes and about inadequacies in our original list of eight. The revised form includes 10 objectives.

2. While student progress on relevant objectives continues to be our criterion of success, the revision more explicitly recognizes that such progress may be a function of the students as well as the teacher. Therefore, a number of items relating to student motivation and expectations have been included and will be examined for their relevance to student progress. These items may help us adjust for the advantage which courses enrolling motivated majors have always had over general education courses.

3. Altering classroom behaviors is one way to induce more student progress; another may be to plan the course more wisely. A set of items on the revision is directed to the latter strategy by inquiring about course demands, content, and reading assignments.

4. To satisfy faculty members and student alike, we dropped the "true-false" format in favor of five-response alternatives throughout. In the course of doing this, non-functioning items (those unrelated to any kind of progress ratings) were eliminated.

5. By using a new input procedure, we have reduced processing costs to an average of $1.50 per class plus 1 cent per student.

We're confident the changes will make the program more valuable to its users.

Our experience suggests that several ingredients are needed to develop a successful program for appraising and improving instructional performance. A sound rationale which can respond meaningfully to well-intentioned concerns and objections of faculty members is essential. The rationale needs solid statistical and research support. Delicate political problems must be faced and resolved with sensitivity, patience, and a willingness to compromise. A smoothly functioning administrative process is essential so that needed materials show up at the proper time and place, no results get lost, reports are made in a reasonable length of time, and continuity in service is assured. Finally, it is necessary to demonstrate sincerity of motive by providing assistance in interpreting diagnostic reports and responding constructively to the shortcomings they identify.

REFERENCES

Bloom, B. S. (Ed.) Taxonomy of Educational Objectives. Handbook I: Cognitive Domain. New York: McKay, 1956.

Deshpande, A. S. & Webb, S. C. Student Perceptions of Instructor Teaching Goals. III. Internal Structure of Ratings. Research Memorandum 68-5, Office of Evaluation Studies, Georgia Institute of Technology, 1968.

Holland, J. L. & Lutz, S. W. The Predictive Validity of a Student's Choice of Vocation. Personnel and Guidance Journal, 1968, 46, 428-434.

Hoyt, D. P. Improving Instruction Through Student Feedback,Manhattan,Kansas: Kansas State University, Office of Educational Research, 1969.

Isaacson, R. L.; McKeachie, W. J.; Milholland, J. E.; Ling, Y. G.; Hofeller, M.; Baerwaldt, J. W.; and Zinn, K. L. Dimensions of Student Evaluations of Teaching. Journal of Educational Psychology, 1964, 1964, 55, 344-351.

Keefer, K. W. Self-Prediction of Academic Achievement by College Students. Unpublished doctoral dissertation, University of Tennessee, 1965.

Krathwohl, D. R.; Bloom, B. S.; & Masia, B. B. Taxonomy of Educational Objectives. Handbook II: Affective Domain. New York: McKay, 1956.

Soloman, D. Teacher Behavior Dimensions, Course Characteristics, and Student Evaluations of Teachers. American Educational Research Journal, 1966, 3, 35-47.

Soloman, D. Rosenberg, L., and Bezdek, W. E. Teacher Behavior and Student Learning. Journal of Educational Psychology, 1968, 59, 297-301.

Walsh, W. B. Validity of Self-Report. Journal of Counseling Psychology, 1967, 14, 18-23.

Whitlock, G. H. The Evaluation of Teaching Effectiveness. Proceedings. Sixth Annual Forum of the Association for Institutional Research, May, 1966.

TEACHING REPORT FOR DR. XXXXXXXXXXX, COURSE XXX-XXX

	SPRING, 1971	SPRING, 1972
PART I. INSTRUCTIONAL METHODS	(BEFORE)	(AFTER)
PREPARATION AND ORGANIZATION	74	96
STUDENT INVOLVEMENT	78	87
CLARITY OF COMMUNICATION	43	93
STIMULATION	67	91
SPEAKING STYLE	69	95
PERSONALISM-CONSIDERATION	59	94
TOTAL	65	93
PART II. PROGRESS RATINGS		
**FACTUAL KNOWLEDGE	3.7	4.4
**PRINCIPLES, THEORIES	3.5	4.2
*APPLICATIONS	3.4	3.9
SELF-UNDERSTANDING	3.1	3.3
PROF. ATTITUDES, BEHAVIOR	3.4	4.0
EFFECTIVE COMMUNICATION	3.2	3.4
*INFLUENCE ON PERSONAL-PROF CONDUCT	3.3	4.0
GENERAL-LIBERAL EDUCATION	2.2	2.6
PROGRESS, RELEVANT GOALS	3.5	4.2
PART III. COURSE RATINGS		
EXAMINATIONS	33	76
ASSIGNMENTS	93	96
TEXTBOOK	82	91
CONTENT	82	90
RECOMMEND TO FRIEND		
AS PROF-COURSE	4.1	4.5
AS PSNL INTRST COURSE	2.0	2.7
INSTRUCTOR	1.6	2.7

Figure 1. Student Ratings of a Single Course and Instructor, Before and After the Instructor Pursued an Improvement Program.

STUDENT EVALUATION OF INSTRUCTION AT MICHIGAN STATE UNIVERSITY

Willard G. Warrington

Michigan State University

It is my intention to follow the case study approach in my presentation, a sort of "show and tell." I want to discuss in some detail the Student Instructional Rating System (SIRS) that has been in operation on the campus of Michigan State University since 1969. The rating form for this system was empirically developed over a two year period, was accepted as a part of the academic program by the faculty and has now been administered in over 10,000 classes to more than 400,000 students.

It is not my intention to argue that our SIRS is the best system in operation anywhere or even that it is an outstanding system but rather to report some of its characteristics and some of the consequences of its widespread use during the past three years.

First, let me put SIRS into some historical perspective. MSU, like many institutions, has for many years encouraged its faculty members to utilize student feedback in analyzing and evaluating classroom instruction. A series of locally developed rating forms were made available but these varied considerably in quality and their use, at best, was relatively infrequent and unsystematic.

Consequently, in 1967, a specific project was funded under the MSU Educational Development Program (EDP) to undertake the systematic development of a comprehensive student instructional evaluation system which would provide faculty members with student reactions to their teaching. This project was under the direction of Dr. F. Craig Johnson, Assistant Director of EDP, (now a professor of Institutional Research at Florida State University) with most of the actual development work being carried out by two doctoral students in Psychology, Wallace Berger and Stanley Cohen.

In view of certain developments which I will discuss later in this paper, it is important to remember that the initial objective of the SIRS project was to develop procedures for allowing an instructor to collect student feedback data that he could utilize for self-examination and self-improvement of his instruction.

In the early stages of this project two important decisions were made that had much to do with the final characteristics of SIRS. First, it was decided to heavily involve both students and faculty in the determination of the content of the rating form that was to be developed and, second, it was agreed that the completed system would provide normative data so that faculty members could determine their standing relative to other faculty teaching similar courses.

No effort will be made to describe in complete detail the actual steps in the two years of the development of SIRS. (For those interested, a fifty page, technical bulletin is available through the MSU Office of Evaluation Services.) However, some broad overview is necessary to understand the system that finally emerged.

Briefly, the SIRS project proceeded as follows: Students and faculty in a wide range of courses were interviewed in the Summer of 1967 using the "critical incident" approach. Faculty were asked to "compare and contrast your best and worst students." Students were asked to "compare and contrast your best and worst instructors." All interviews were content analyzed resulting in 1300 key phrases and sentences which were rewritten in an item format suitable for an evaluation form. Items from existing student instructional rating forms were also collected. After much editing and elimination of duplication, 223 experimental items were grouped into six parallel forms for administration during the 1967, Fall Term.

A rather elaborate stratified sampling procedure identified 1,286 students and 594 faculty members who were asked to react to one of the parallel forms by responding to four questions about each item. These questions represented evaluative dimensions which emerged from the initial interviewing. The faculty was asked:

1. Does this item present information which you could use for course improvement? (yes/no)
2. If you were to construct a student course appraisal sheet would you include this item? (yes/no)
3. Would you need additional information to interpret the responses to this item? (yes/no)
4. Do you believe that students have enough information and/or are competent to accurately respond to this item? (yes/no)

Students answered the following questions for each item:

1. Do you believe this item is relevant for appraising this course? (yes/no)
2. If you were to construct a student course appraisal sheet would you include this item? (yes/no)
3. Would you want to qualify your response to this item? (yes/no)
4. Do you believe that you have enough information and/or are competent to evaluate those aspects of the course referred to by this item? (yes/no)

Of the 1,286 questionnaires mailed to students, 611 returns were usable for a return rate of 48%. Of the 594 faculty questionnaires mailed out, 265 of the returns were usable, for a return rate of 45%.

The data were analyzed in the following manner: The proportion of the students indicating that an item was (1) relevant for evaluation of their course, (2) potentially useful in terms of course improvement, (3) not needing student qualification, and (4) capable of meaningful student evaluation, was computed. The same was done for faculty responses. The students had also been asked to evaluate their course through the use of the experimental items. These responses were used in order to determine the response distribution on the items.

The intercorrelations among the four questions were computed separately for faculty and students. The two (faculty and student) intercorrelation matrices showed some similarity and some striking differences. For the faculty the correlation between whether an item could be used for course improvement and whether the item should be included in a course evaluation form was .95. For students the correlation between whether an item was relevant for course evaluation and whether it should be included in the evaluation form was .96. Even though there was a strong relationship between the inclusion of an item in an evaluation form and its usefulness or relevance for both faculty and students, there was not nearly as high agreement (r = .68) between faculty and students as to whether an item should be included. This undoubtedly accounts for some of the disagreements between faculty and students as what should be on an instructional rating scale.

Furthermore, for the faculty, negative correlations were found between question 3 ("Would you need additional information to interpret the response to this item?") and the other three questions. For the students, question 3 ("Would you want to qualify your response to this item?")

yielded no negative correlations with the other three questions. Thus, it seemed that faculty wanted additional information with respect to many items but students felt that a categorical yes-no answer was sufficient. And another source of difficulty may result from the fact that the correlation between whether the faculty believe that the students have enough information to answer the item and whether the students believe that they have enough information was relatively low, only .54. Very likely, this difference of opinion between students and faculty decreases the effectiveness of student involvement in educational decision-making, in general.

As the study proceeded it was agreed that a subset of the original 223 items would be selected for a second experimental SIRS form by equally weighting faculty and student opinion in the following manner.

An item was selected for inclusion in the next experimental form if:

1. At least 70 percent of the students and at least 70 percent of the faculty indicated that the item (a) could be used for course improvement (is relevant for course appraisal), (b) should be included in an evaluation form, and (c) could be competently evaluated by students.
2. It had a pooled student and faculty average higher than 80 percent on the above three variables.

If any of the items which fulfilled these two criteria were rated by 40 percent or more of the students or the faculty as needing qualification, the items were rewritten and then included in the form.

Thus, only those items which were judged as being relevant, warranting inclusion in an evaluation form, and capable of meaningful student evaluation by 80% of the combined student and faculty sample were designated as pilot items.

Through the use of the above procedures 56 items were selected for the first pilot form. These items were divided into six categories, each category preceded by a title which appeared to characterize the general topic covered. Eleven biographical items, (class level, age, course required, sex, course recommended, marital status, number of credits earned, preconceptions of this course, G.P.A., number of other courses in the same department, and grade up to now) were also included in the questionnaire in order to assess the relationship, if any, among these variables and the course evaluation items.

This pilot form was administered in the winter of 1968 to 2,841 students in large introductory level courses taught by 36 different instructors. Various types of analyses were performed on the resulting data including a Varimax factor analysis. Five factors were identified and interpreted as follows:

Factor 1. Consisted of eight items and appeared to be related to instructor characteristics such as instructor involvement and attitude towards teaching. (INSTRUCTOR INVOLVEMENT)

Factor 2. Consisted of seven items and appeared to be related to the students' interest in the course and the students' performance in the course. (STUDENT INTEREST AND PERFORMANCE)

Factor 3. Consisted of six items and appeared to be related to student-instructor interaction in terms of personal communication between students and faculty members. (STUDENT-INSTRUCTOR INTERACTION)

Factor 4. Consisted of five items and appeared to be related to the difficulty and speed at which the course material was presented. (COURSE DEMANDS)

Factor 5. Consisted of seven items and appeared to be related to the organization of course materials and lecture presentations. (COURSE ORGANIZATION)

From data collected thus far a 20-item scale consisting of four items for each of the five factors was tested in the summer of 1968 to (1) determine the stability of the factor structure under both a two-choice and a five-choice item format and (2) pre-test a machine-scored form. Approximately half of the 1,200 students tested received the two-choice form and the remainder the five-choice form. These data indicated that the factor structure was stable and that the five-choice format had superior operating characteristics in addition to being more favorably received by both students and faculty.

During the remainder of 1968, items pertaining to laboratory and recitation sections were developed by a process similar to the one described above. Also, a general purpose item, No. 21, "you generally enjoyed going to class" was identified and included in the final form which now contained 21 instructional evaluation items, four student background items, and three laboratory and recitation items.

Copies of the instrument, printed on Optical Scanning sheets as attached, were made available to the College of Agriculture, College of Engineering, College of Social Science and the University College in the spring of 1969. These four colleges administered and had scored 8,012 forms.

For a last comprehensive check on the item structure, correlations over the 21 instructional evaluation items for the 8,012 respondents were computed, factor analyzed, and subjected to a Varimax rotation. The structure remained stable and the pattern of item loadings was identical to that of earlier studies.

It, therefore, seemed reasonable to assume that feedback to the instructor could consist not only of the mean responses on the 21 items but also of the mean response of each of the five factors. Since the loadings for each item were of a high magnitude, a good approximation to the factor means could be obtained by simply averaging the means of the four items most heavily weighted on each factor. The five averages would comprise a composite profile of the instructor's evaluation on the five dimensions of the learning situation. This format was later incorporated into the SIRS Report. Internal consistencies (average inter-item correlations corrected by the Spearman-Brown formula) were computed for each of the five dimensions and were as follows:

Instructor Involvement - .81
Student Interest and Performance -.79
Student-Instructor Interaction - .84
Course Demands - .73
Course Organization - .83

After the analysis of the data from the spring 1969 administration, the decision was made to consider the rating form finalized and to proceed to develop the rest of the evaluation system. It is important to remember that SIRS was never seen simply as a paper-and-pencil rating instrument but rather as a system for collecting, analyzing, displaying and interpreting student reactions to classroom instruction and course content in order to improve the quality of that learning situation. The rating form obviously related to the collecting aspect of the system. It should be noted that the final form contained some blank spaces in which the instructor may insert optional items of his own choosing. The student responses for these items would be summarized in the SIRS Report which is discussed below. In addition to this flexibility, the back of the SIRS Form is available for more general comments or specific reactions to specific questions.

The SIRS Form was designed to be processed by an Opt Scan 100 Dm Optical Scanner which produces a 800 cpi-9 channel magnetic tape. This tape is read into an IBM 370-155 (initially an IBM 360-60) computer which analyzes the student responses and produces a print-out known as the SIRS Report. Much trial and error went into the programming and format of this Report to make it readable and understandable by faculty members with little or no computer experience.

The Report lists each question from the SIRS Form along with the percentages of students marking each of the five positions from Strongly Agree to Strongly Disagree. The mean and standard deviation of responses for each question are also shown. These are computed using a 5-point scale where Strongly Agree is assigned a value of 1 and Strongly Disagree a value of 5. Also for each question, percentile ranks are given indicating how the mean for this particular administration compares with previous administrations of SIRS in the same course, in all courses in that particular department and in all courses in that particular college. In all cases the percentile rank listed indicates the percent of previous administrations that resulted in mean ratings that were less favorable than the present mean rating. In other words a high percentile rank indicates that the mean rating for a question in this particular administration is higher than most mean ratings from previous administrations.

Of course, this audience recognizes that this is a relative system of comparison and that in any given situation half of the administrations will result in mean ratings that will be above average and half below average, regardless of the general level of instruction. Nevertheless, in our opinion, it seems desirable to present this comparative data since otherwise student reactions are very hard to interpret because they tend to be overly positive. Such inflated ratings often present a misleading picture to the instructor who receives mean ratings near or above the midpoint of the scale where, in fact, his ratings may be quite low when compared with many of his colleagues.

In addition to the data presented for each question the Report also presents a composite Profile in which summarized data is shown for each of the five areas mentioned above, namely Instructor Involvement, Student Interest, Student-Instructor Interaction, Course Demands, and Course Organization. Here again means, standard deviations and percentile ranks for the course, department and college are shown for each area. Since these data are based upon the average over four questions for each area, they tend to be somewhat more stable than data for individual questions.

To assist users of SIRS in understanding and interpreting their Report a SIRS Manual was developed to support the system. The Manual summarizes the purposes and characteristics of SIRS, gives information as to what this data mean and how they should be interpreted, lists some precautions in using the Report and presents a variety of questions that instructors may want to use as optional items when they administer the SIRS Form. To date, several thousand of these Manuals have been distributed and the overall reaction to the document has been quite favorable.

After the total system had been thoroughly reviewed and experimental administrations had been given to another several thousand students in the fall of 1969, SIRS was recommended for adoption on a mandatory basis for the total University. After considerable debate, all generally constructive, the University Academic Council, which is the highest faculty governance entity, on December 2, 1969, passed the following resolution:

> Use of the Student Instructional Rating Report
>
> The use of the Student Instructional Rating Report (SIRR) should be adopted with the full realization that it is but one parameter of instructional evaluation.

A. The regulations for the use of Student Instructional Rating Reports in effect since January 20, 1949, will be declared void on adoption of the new policy.

B. Each of the teaching faculty (including graduate assistants) at Michigan State University regardless of rank or tenure is required to use the Student Instructional Rating Report to evaluate (1) at least one course in every quarter in which he teaches and (2) every separate course he teaches at least once a year.

C. The results generated by the Instructional Rating Report shall be evaluated at the departmental level in order to help determine individual effectiveness. Appropriate procedures for the execution of this evaluation shall be determined according to departmental or residential faculty perrogatives.

Two aspects of this action seemed rather interesting. First, the resolution made the use of SIRS mandatory across the board. The requirement that an instructor obtain student feedback pertaining to his instruction no longer applied only to lower ranks and/or to relatively new faculty members as had previously been true. But second, and even more drastic, the resolution for the first time officially recognized that student reactions to instructors are no longer the sole property of the particular faculty member, but belong, in part, to that segment of the University involved in making decisions with respect to the academic effectiveness of this faculty member. I am still not completely convinced that the Academic Council members were fully aware of the implications of the resolution they passed and now, over three years later, it has not been seriously challenged.

I would like to discuss briefly one additional aspect, a very important one in my opinion, of SIRS before reporting what has happened since the system was adopted officially. Any system must be internally reinforcing if it is to be self-improving. The components of SIRS described above will be internally reinforcing to the extent that the Form is accepted as useful and the Report is relevant and understandable. But there must be an additional mechanism to make the system complete. To be specific, if an instructor who is utilizing the system decides, from his results, that he needs additional assistance, such assistance must be available. At MSU, the Instructional Development Services in the Provost's Office is designed to serve this function. The Instructional Development Service includes three different supportative agencies: (1) The Learning Services assists instructors in analyzing their instructional situations, in the development of their objectives and in the structuring of actual learning experiences. (2) The Instructional Media Center provides a full range of consultative and supportive services in the audio-visual area, including closed circuit television, and (3) The Office of Evaluation Services provides consultation services and technical assistance in the area of classroom evaluation and test construction and analysis. All three of these offices have well qualified professionals who work in a face-to-face situation directly with faculty members who are trying to better understand and improve the learning that takes place in their classrooms. This is the segment of an instructional evaluation system that is too often lacking. Granted, these are relatively expensive operations but, in my opinion, they are vital if the quality of instruction is to be improved through the utilization of student evaluations of instruction.

Now back to a brief discussion of what has happened since December, 1969, with respect to SIRS. First, the system is being used widely. As of the middle of fall term, 1972, normative data was available for 9,326 administrations involving 318,654 student responses. Nearly 100,000 more responses have been collected since then. SIRS administrations have been processed for classes in every college and most teaching departments of the University.

However, it should be made clear that not all departments are reacting the same to the December 2, 1969, Academic Council Resolution which you recall did give the department considerable leeway as to how SIRS should be utilized. Several departments are stressing the requirement that all classes be evaluated and that the results must be available at the department level. Others are allowing the individual instructor to decide whether or not he submits his SIRS results to the department. And a few departments have decided that the SIRS form is not appropriate and have developed an instrument of their own. Several of these incorporate much of the SIRS approach, others are completely different.

Comments with respect to SIRS have ranged from very supportive to very critical. The most common criticism is that the Form is too "blah," i.e., it does not ask the important questions. This comes from both students and faculty. But if you recall the method by which the questions were selected this is not entirely unexpected since the original data indicated considerable disagreement between faculty and students as to what was important and what should be included on an appraisal form. Yet, only those items upon which there was high agreement were included. We recommend that instructors include those questions about which they feel strongly as optional items on the SIRS Form or administer a complementary form in addition to the SIRS Form.

Another area of concern has developed which is much more difficult to cope with. Many faculty members are quite concerned with the lack of uniformity as to how the SIRS forms are administered and used. These people feel that if faculty personnel decisions are to be based, even in small part, on SIRS results, then it is important that such data are collected under the same standardized procedures. It is our belief that systems for administering the SIRS forms should probably be the responsibility of colleges or departments. A university-wide system would probably be unwieldy and unresponsive to departmental needs.

Two other areas of concern are worth mentioning. There seems to be some tendency on the part of some faculty and administrators to act as if we have the problem of the evaluation of instruction solved. Many of us have argued strongly but evidently not too effectively that student evaluation of instruction is only one dimension of this overall evaluation process. In our opinion, classroom visitation and observation by colleagues and administrators can provide useful data. Similarly, some of us would like to see more attention given to attempts to measure changes in student behavior as relevant information for evaluating teaching effectiveness. However, I do not want to minimize the importance of the student input but only to argue for additional systematic input into the total process.

And finally we are genuinely concerned about improper or over-interpretation of the data provided in the SIRS Report. We occasionally hear where some instructor is called into question because his comparative norms have dropped a couple percentage points. Or some department chairman cannot understand why some people in his department are below average. Or some instructor receives a mean rating considerably above the midpoint of the scale yet receives a normative rating at the 30 percentile rank. We try to answer these queries by phone or in person when they are brought to our attention. But in an attempt to answer these and other unasked questions we have to date prepared four SIRS Research reports.[1] These are:

---

[1] These reports and other SIRS support materials are available from the Office of Evaluation Services, 202 South Kedzie Hall, Michigan State University, East Lansing, Michigan 48823.

1. Analysis of SIRS Responses for Winter Term, 1970 - Feb., 1971
2. Stability of Factor Structure of SIRS - Nov., 1971
3. Using SIRS Data in the Decision-Making Process - March, 1972
4. Student Instructional Rating System Responses and Student Characteristics - May, 1972

Very briefly, Report #1 presents summary data from early SIRS administrations for the total University (remember our norms are for course, department and college). Data is also given for SIRS responses by level of course, by reason for taking course, and by level of grade point average.

Report #2 presents the two independent SIRS factor structures that were produced as the instrument was developed. It is interesting to note that Dr. Raoul Arreola of Florida State University reported at the 1973 meeting of NCME in New Orleans that he had factor analyzed the results of the MSU SIRS administrations at FSU and had gotten a factor structure very similar to that which we had obtained. His data further supports the rather remarkable stability of the factors mentioned earlier in this paper.

Report #3 was designed to provide SIRS users, particularly those using SIRS data in personnel decision-making, with a more sophisticated explanation of the nature and limitations of SIRS data, especially the percentile norms. Precautions and illustrations of appropriate and inappropriate interpretations were discussed in considerable detail. We have some evidence that this document has been useful but it has certainly not eliminated all problems in the area of utilization and interpretation of SIRS data.

Report #4 is the first of what we expect to be a series of more specific research oriented presentations. This particular study investigated the effect of administrating SIRS forms under two different conditions of student identification. One, the regular condition of anonymity and a second mode in which the student records his student number on the SIRS form.

The latter method of administering the SIRS form would make it easier to design studies to investigate relationships between student characteristics and responses to the SIRS form. While the results of the study are somewhat limited, there is considerable evidence that the change from student anonymity does change the SIRS responses. This suggests that it will probably be necessary to collect student characteristic data in the same anonymous fashion and at the same time as the SIRS administration if these interrelationships are to be meaningfully studied.

Another SIRS study presently underway investigates the type of response scale. The SIRS form uses Strongly Agree to Strongly Disagree. Students tend to use only the Strongly Agree or Agree response categories for many SIRS items. While it is gratifying to know that MSU students have such positive attitudes toward their instructors, it is difficult to make statistically meaningful discriminations between instructors. One of our graduate students is conducting a doctoral study of alternative response scales to see if student responses can be made less lenient and, therefore, more discriminating.

But what of the future of SIRS at MSU? Certainly all is not sweetness and light so SIRS will continue to receive more than its share of scrutiny due to the delicate area with which it is concerned. The use of the same instructional rating form for both administration decision-making and as a feedback mechanism to the instructor for purposes of improvement will continue to be questioned. We are inclined to think that it would be better to have two types of instruments to meet these quite disparate purposes. It might be better to use one form that concentrates on widely-accepted instructional practices such as meeting the class regularly, clearly defining the objectives of the course, communicating to students the methods of student evaluation, and so on. This "responsibilities" form could be systematically administered by the departments and the results used in making faculty personnel decisions. In addition, instructors could use the present SIRS form or an extended version specifically tailored to specific instructional settings to provide them with diagnostic data that would be more useful for instructional improvement. Results from this latter type of feedback could be submitted through departmental channels if the instructor so desired but would not be required.

Some changes probably need to be made in the SIRS norm system. It might be better to report percentile bands rather than specific percentiles since the present system suggests a higher degree of precision than we would prefer. The question of current norms vs. cumulative norms also needs further attention. Cumulative norms, as the system presently uses, maximizes sample size which is important in courses and departments with small enrollments. However, attitudes of students do change markedly over time which reduces the value of data obtained some time earlier. Very likely some combination of current norms, say from one term earlier, for large enrollment areas will be introduced.

A subcommittee of our University Educational Policies Committee, the committee that approved the original recommendation to our Academic Council, has been assigned the task of reviewing the present status of SIRS and making recommendations for its improvement. Most of the points discussed in this paper have been brought to that group's attention. While it is unwise to predict the outcome of a committee's deliberations, I expect that, while some changes will undoubtedly be recommended, the present Student Instructional Rating System will continue to be a viable aspect of the instructional program of Michigan State University.

BEST COPY AVAILABLE

Appendix

MICHIGAN STATE UNIVERSITY

STUDENT INSTRUCTIONAL RATING SYSTEM FORM

SA if you strongly agree with the statement
A if you agree with the statement
N if you neither agree nor disagree
D if you disagree with the statement
SD if you strongly disagree with the statement

[illegible] the KEY

KEY

		SA	A	N	D	SD
1. The instructor was enthusiastic when presenting course material.	1.	SA	A	N	D	SD
2. The instructor seemed to be interested in teaching.	2.	SA	A	N	D	SD
3. The instructor's use of examples or personal experiences helped to get points across in class.	3.	SA	A	N	D	SD
4. The instructor seemed to be concerned with whether the students learned the material.	4.	SA	A	N	D	SD
5. You were interested in learning the course material.	5.	SA	A	N	D	SD
6. You were generally attentive in class.	6.	SA	A	N	D	SD
7. You felt that this course challenged you intellectually.	7.	SA	A	N	D	SD
8. You have become more competent in this area due to this course.	8.	SA	A	N	D	SD
9. The instructor encouraged students to express opinions.	9.	SA	A	N	D	SD
10. The instructor appeared receptive to new ideas and others' viewpoints.	10.	SA	A	N	D	SD
11. The student had an opportunity to ask questions.	11.	SA	A	N	D	SD
12. The instructor generally stimulated class discussion.	12.	SA	A	N	D	SD
13. The instructor attempted to cover too much material.	13.	SA	A	N	D	SD
14. The instructor generally presented the material too rapidly.	14.	SA	A	N	D	SD
15. The homework assignments were too time consuming relative to their contribution to your understanding of the course material.	15.	SA	A	N	D	SD
16. You generally found the coverage of topics in the assigned readings too difficult.	16.	SA	A	N	D	SD
17. The instructor appeared to relate the course concepts in a systematic manner.	17.	SA	A	N	D	SD
18. The course was well organized.	18.	SA	A	N	D	SD
19. The instructor's class presentations made for easy note taking.	19.	SA	A	N	D	SD
20. The direction of the course was adequately outlined.	20.	SA	A	N	D	SD
21. You generally enjoyed going to class.	21.	SA	A	N	D	SD
22.	22.	SA	A	N	D	SD
23. Instructor may insert three (3) items in these spaces.	23.	SA	A	N	D	SD
24.	24.	SA	A	N	D	SD

STUDENT BACKGROUND Select the most appropriate alternative.

25. Was this course required in your degree program?	25.	yes	no			
26. Was this course recommended to you by another student?	26.	yes	no			
27. What is your overall GPA? (a) 1.9 or less (b) 2.0 - 2.2 (c) 2.3 - 2.7 (d) 2.8 - 3.3 (e) 3.4 - 4.5	27.	a	b	c	d	e
28. How many other courses have you had in this department? (a) none (b) 1 - 2 (c) 3 - 4 (d) 5 - 6 (e) 7 or more	28.	a	b	c	d	e
29. Instructor may insert two (2) items in this space.	29.	a	b	c	d	e
30.	30.	a	b	c	d	e

DO NOT WRITE BELOW THIS LINE UNLESS THIS COURSE HAS LABORATORY OR RECITATION SECTIONS

LABORATORY or RECITATION (fill in your recitation or lab number at the bottom)

31. The laboratory or recitation instructor clarified lecture material.	31.	SA	A	N	D	SD
32. The laboratory or recitation instructor adequately prepared you for the material covered in his section.	32.	SA	A	N	D	SD
33. You generally found the laboratories or recitations interesting.	33.	SA	A	N	D	SD
34. Instructor may insert two (2) items in this space.	34.	SA	A	N	D	SD
35.	35.	SA	A	N	D	SD

IMPORTANT

WRITE and **MARK** in the boxes to the right your recitation or laboratory section number. Section number 1 would be written and marked 001, section number 15 would be written and marked 015. If you do not have a recitation or lab section leave this area blank.

RECITATION OR LABORATORY SECTION NUMBER

1		0	1	2	3	4	5	6	7	8	9
2		0	1	2	3	4	5	6	7	8	9
3		0	1	2	3	4	5	6	7	8	9

Michigan State University Printing Service [illegible]

BEST COPY AVAILABLE

Appendix (continued)

STUDENT INSTRUCTIONAL RATING SYSTEM FORM (Written Comments)

[illegible] instructor can improve his class is through thoughtful student reactions. This instructor [illegible] responses for self-examination and self-improvement. If you have any comments to make about the instructor or the course, please write them in the shaded area below.

CRITERIA FOR THE EVALUATION OF COLLEGE TEACHING: THEIR RELIABILITY AND VALIDITY AT THE UNIVERSITY OF TOLEDO

Richard R. Perry and Reemt R. Baumann

University of Toledo

This paper contains two complementary sections. The first section, written by R. Perry, describes the search for teaching characteristics which are critical in the discussion of teaching effectiveness. R. Baumann provides information about the Student Perception of Teaching Effectiveness Scale which was built primarily upon the findings of R. Perry's study, and has been used for six years at the University of Toledo, College of Education.

Introduction

Faculty of colleges and universities have always been under the searching eye of those who evaluate performance. This evaluation is prompted, hopefully, by the widespread interest of society in the educational process. Widespread interest and consequent evaluation has sometimes had serious effects on those who are being evaluated. We are all aware that Socrates was executed in Athens in 399 B.C. as a result of the evaluation of his teaching which ended with the accusation that he should be done away with because of "introducing new gods and corrupting the youth." We are aware that in the early medieval universities physical abuse and death was sometimes the consequence of the evaluation of teaching. Cecco de Sacoli was burned at the stake at the University of Padua in 1237 for ineffectiveness in his teaching of astrology. George Whitfield, member of the faculty at Harvard in 1745, was severely censored for being impious and enthusiastic and possessing a conceit about his own worth and excellence. This all resulted because he had published a paper in which he accused the universities for having now "become darkness--darkness that may be felt where previously teaching. Some of those evaluations result in accolades others have different effects.

The Problem

It seems that one of the major difficulties associated with evaluation is that if evaluation is going to take place, someone, somehow, must identify the criteria on which the evaluation could be based. That has been a major problem in evaluation of college teaching.

Identification of teacher effectiveness is so complex that apparently no one knows today what "the competent teacher is." The anonymity of the "competent teacher" has been the spur for countless research studies. Gage (1960) stated that literature on teacher competence is overwhelming; so much so that even bibliographies on the subject are unmanageable. Although numerous studies are reported in the literature, few if any "facts" are firmly established about teacher effectiveness. There is no approved method of measuring competence which has received wide acceptance (Biddle, 1965). The statements by Gage and Biddle support the need to focus attention on the identification of criteria. Harm that can be accomplished by using inappropriate criteria suggests research to identify characteristics of effective teaching behavior.

One of the most serious aspects of the problem of identifying effective teaching behavior is that without such explicit identification evaluations which take place are suspect. Significant faults which are assigned to present methods of evaluation focus chiefly on the following inadequacies:

1. Criteria included in evaluations have not been warranted by adequate research.

2. Persons who do evaluation are criticized for their lack of expertness in the very field in which they are operating.

3. Evaluation of teaching behavior has not proven to produce high reliability in longitudinal studies, when total effectiveness of teaching behavior is considered.

The lack of conclusiveness of previous investigations has not diminished the zeal with which the results of such investigations are put forward. Perhaps, the most useful result of all such examinations and experiments is to more clearly identify the problems experienced in trying to arrive at clear definitions of effective teaching. A most important consideration in such research is to understand that substantive evaluation can take place only in terms of explicit objectives. Until objectives are defined and agreed upon evaluations tend toward spuriousness. However, a corollary to the establishment of objectives is the identification of criteria of teaching behavior which, hopefully, will elicit, or at least assist in, the attainment of teaching objectives. Even when a careful definition of desirable outcomes (objectives in teaching) is attained, it does not solve the criterion problem.

After objectives have been established for an educational program, it is necessary to identify those criterion behaviors which will have to produce the objectives, the criterion behaviors of teaching related to the objectives are then useful in the pursuit of evaluation of the teaching. Since the major problem in research on teaching behavior is that of criteria (McKeachie, 1963), it seems that research on the identification of criteria which can be warranted for the evaluation of effective teaching behavior might be helpful.

Such attempts in higher education are not new. They have increased in frequency in the last ten years. Research on the identification of warranted criteria received much impetus from the work of Ryans whose argument for such research indicates that there are good teachers and good teaching, and that characteristic behavior associated with such teaching should be able to be identified. Even though they may be identified it can be assumed that not every teacher can possess all the "good" behaviors or characteristics; thus the goal of such research needs to be the identification of those criteria of teaching behavior which are critical. The identification of such criteria has been left often to the expert opinion or to administrative standards. The use of such authority has resulted in criteria proving unfruitful and of temporary value. The argument has gained weight that the place to look for characteristics of teaching behavior which result in effective teaching is in the behavior of teachers. Such reasoning suggests searching out clusters of behaviors associated with effective teaching.

A word needs to be said about the meaning of effectiveness. A single piece of research cannot hope to explore all the dimensions implicit in a concept such as effective teaching behavior. The majority of research studies in this area have focused on the assumption that in searching for teaching effectiveness, the research seeks for properties of the teacher. This assumes that effectiveness is an attribute of the teacher. A further assumption is that such effectiveness is not seriously deterred by other variables. This establishes an hypothesis about the adaptability of a teacher to teaching situations (Fattu, 1963).

The assumption that the effective teacher is one who can accomplish educational objectives with students, aside from other variables, is to recognize that effectiveness as a term may have several meanings as it is identified with several different teaching situations. There is no harm in using the term effectiveness as long as it is recognized that it is related to a set of particular conditions.

Effectiveness in teaching in the sense of the study done at the University of Toledo and replicated at the University of New Mexico, Las Cruses, Northern Illinois University and Western Kentucky University was taken to mean those behaviors identified by faculty, students, and alumni which when made operational would result in effective teaching.

A Brief Appraisal of Evaluation of Teaching Behavior

Evaluation of teaching seems to enjoy great attention in the popular and professional press but one needs to remember that systems of such evaluation have been operative in colleges at least since the early 1920's. Some procedures have resulted in evaluations being given to deans or department chairmen who, in turn, are privileged to confer with faculty about the evaluations. Apparently, other systems of evaluation make it possible for the results of such procedures to be made known to salary and promotion committees and others merely have the results made known to the professor.

It seems that none of these systems of evaluation is without criticism and a few of these criticisms are helpful in the identification of basic faults in such evaluations. Major criticisms which are a matter of record in the minutes of faculty meetings at a private college indicate that:

1. The present procedure cannot be intelligently considered as evaluation of effective teaching but would be better named "poll of student opinion."

2. The present system does little to help in determining which faculty will be kept and which faculty will be lost, or which faculty will be attracted to the campus.

3. Those involved in the evaluation are not by education, experience or responsibility qualified to make the evaluation they are asked to make.

4. The fact that the current evaluation is obligatory upon the faculty member is a violation of faculty rights (Antioch College, April 25, 1964).

The above comments represent a core of a faculty's concern about evaluation procedures.

There are other thoughts which are based on inadequacies in systems of evaluation. These seem to center on the following:

1. An institution will decide to provide for evaluation of teaching and will choose evaluation items from rating instruments which are already in use at other institutions.

2. An institution or indeed an entire state system of higher education will decide to honor outstanding teachers with cash prizes but will leave the identification of these outstanding teachers to the judgments of persons in positions of administrative authority, or to impressionistic evaluations of individual faculty. The comment of one professor who found himself involved in a system of higher education providing for such identification indicated that, "even if you wanted to try out for an award you wouldn't know how to change your teaching. This whole reward set-up is too much like a beauty contest (Old Oregon, January-February, 1966, p. 13)."

3. An institution will make it possible for the evaluation of teaching to go on in one college or in one department and not in all of the departments or colleges on a campus. Thus, some faculty feel imposed upon while other feel deprived of the opportunity for evaluations.

4. The most serious concerns about the evaluation of teaching focus on the question, "Evaluation for what purpose?" This question has not been satisfactorily answered on a majority of campuses.

5. An additional area of major concern is finding a satisfactory answer to the question, 'What criteria can be justified in the evaluation of a faculty member's effectiveness as a teacher?"

There is little question but what evaluation of a faculty member's effectiveness as a teacher takes place. Students, his faculty colleagues, and the administration, if he should happen to be known to the administration, all comment in one way or another about the qualities of teaching exhibited by the faculty member. Implicit in all such evaluations is the concept that some faculty must be exhibiting behaviors in their teaching which are considered to be characteristic of effective teaching. Finding out what those behaviors are and determining a relative importance for each of the identified behaviors could be a first step in construction of a model or set of behaviors associated with effective teaching in higher education at any institution of higher education.

The University of Toledo's study on criteria of effective teaching centered on identifying effective teaching behaviors and determining their relative importance.

There are numerous studies which produce interesting statistical results concerning reliability, correlations, and the results of factor analysis. Difficulties in some of these arise because of methods used in selecting criteria for evaluation instruments. Procedures which have established evaluation instruments by choosing criteria already in use at other institutions without testing the warrantability of these criteria for the institution where they are to be used leaves something to be desired. Statistical analysis can be accomplished with responses given to any criteria utilized in any rating instrument, but the question remains as to the warrantability of criteria which are put to use in such procedures.

The University of Toledo Study

Background

Interest in effective teaching is not new to the University of Toledo, but in the last two years it has received increasing attention from the

University faculty and student body. The administration of the University in the Spring of 1964 announced the establishment of four outstanding teaching awards in the amount of $1,000 each. These, financed by the Alumni Foundation are given to four faculty each year in recognition of outstanding accomplishments in teaching at the University of Toledo. The College of Education at the same time introduced structured evaluation procedures for its own faculty. The College of Education provided that at the end of each term faculty members could voluntarily request students to respond to an evaluation instrument which focused on the qualities of teaching in those courses taught by the individual professor. The evaluation instrument not only operated for the individual instructor but for the course as well. The criteria in the instrument resulted from the studied deliberations of a faculty committee of the College of Education. Since 1968, results of the College of Education evaluation procedure were made known to the individual faculty member and to the salary and promotion committee of the College.

The Office of Institutional Research at the University simultaneously with these developments evidenced an interest in conducting a research study within the University community to get at the identification of these criterion behaviors which could be warranted for use in the evaluation of effective teaching behavior in higher education.

The study was proposed to the deans of the colleges and the Faculty Conference Committee, all of whom endorsed it. An advisory committee to the Office of Institutional Research was appointed. The advisory committee consisted of a representative of each college appointed by the dean of that college. The proposed research focused on the central problem of evaluating effective teaching in higher education. That problem without question is the identification of criteria warranted for use in such evaluations, for unless criteria used in such evaluations can be demonstrated as warranted for the purpose at hand, they would be irrelevant.

In structuring the study, the Office of Institutional Research made the following assumptions:

1. Criteria for the evaluation of effective teaching are related directly to the academic community in which they are to be used, and the place to look for these criteria, which are most appropriate for one institution, is within the academic community represented by that institution.

2. Criteria for the evaluation of effective teaching in higher education should be established as the result of consultation with those most directly concerned with such teaching; namely, students, faculty, and alumni of all the colleges of that institution.

3. Students, faculty, and alumni should have opportunity to express their thoughts freely as to what separate actions they believe contribute to effective teaching, without their responses being limited by procedures which force them to select behaviors from a suggested list of such criteria which do not originate within their own community.

The First Phase

The University of Toledo began in the Spring of 1965 and proceeded during the academic year 1965-1966, with the first phase of the study, with the second phase being completed in the academic year 1966-67. The first phase contacted a stratified sample of faculty, students, and alumni to obtain free response identifications of behavior which contributed, in the judgment of the respondents, to the effectiveness of teaching. In order that this could be done and the data handled effectively, response instruments were designed to the configuration of a data card. The response instrument, along with a personal data card, was mailed to a random sample of the student body stratified by college and class rank, to every member of the faculty of the University of Toledo, and to a random sample of alumni stratified by college from which they had received their degrees. Each potential respondent of the sample was given a personal data card and fifteen response instruments.

Thirteen thousand six hundred and forty-three (13,643) individual responses were received identifying "effective teaching behaviors." These responses were received from 812 students, 166 faculty, and 665 alumni. This resulted in replies from 10% of the student body, 30% of the faculty, and 8% of the alumni degree holders. The mean of behaviors identified by students was 8.7. The mean by faculty, 8.2; the mean from alumni, 6.8.

These 13,643 identified behaviors were then "read" by a jury group to identify duplications in behaviors. The jury group was looking for criterion statements which said the same thing essentially, although the wording of the criterion behavior statement might have been different. Examples are the two following responses:

1. "Ability to keep presentation of subject matter at a level comprehended by the student."
2. "Ability to present subject matter at student level."

Though the wording is slightly different in each statement, each can be valued as meaning the same as the other. The result of this reading process was to categorize 13,643 individual behaviors into 60 criterion behaviors. The reading procedure had one jury person read the statements placing them in categories of sameness and then to have these categories checked by second and third jury persons; thus, questions were raised as to the appropriateness of the classification of any one of the criterion statements.

An additional result of this reading process was to identify six major categories of effective teaching behaviors. These six categories contained individual behaviors which grouped themselves into major behavior categories representing concentrations of similar kinds of behavior so as to permit their identification as major separate areas of teaching behavior. The identification of the individual criterion behavior and the clustering of these into the six major criterion behavior areas ended the first phase of the study.

The Second Phase

With the criterion statements on hand, the task was to obtain judgments of how warranted these were for the evaluation of effective teaching behavior. This was accomplished by designing a response instrument in which the criterion behaviors were listed. The order of their listing was provided by a random listing of numbers supplied by a random number program from the University computer. The instruments provided for a response to the importance of each criterion from critical importance through no importance. Each respondent was able to categorize himself by checking appropriate spaces.

A sample of students stratified by college and class rank and a similar sample of alumni by college in which they had earned degrees was presented with the instrument along with all faculty. Usable responses were returned by 756 students, 850 alumni, and 187 faculty. Returns resulted in replies from 7.5% of the students, 8.6% of the degree holding alumni, and 35% of the faculty. These percentages of the academic community seemed adequate in view of present research practices (Holland and Richards, 1965). Weights of 5, 4, 3, 2, and 1, respectively, were assigned to the response areas of critical, above average, average, below average, and no importance. These data were coded into punched cards and processed for statistical analysis to establish rank order correlations for selected categories of responses. Of the 82 rank order correlations calculated, 40 were in the .90's, 34 in the .80's, and 8 in the .70's, all well beyond the .01 level of significance.

The Third Phase

The University of Toledo identified four outstanding teachers in each of the years 1964, 1965, and 1966. Responses of this group were obtained and processed for the same statistical analysis as for other selected respondent categories. The rank order correlations between the "Outstanding Teacher" group and other groups were all greater than .70, well beyond the .01 level of significance. The correlations of the ranking of the criteria by the outstanding teachers with those of all other groups in the study has the effect of testing the order of importance established in the study against the judgments of a "jury of experts." Seemingly, this is further justification for the warrantability of the criteria in the order established for them by the responses of the total group.

A Possible Weighting Procedure

A criticism often leveled at evaluation procedures is that each criterion is assumed to be of the same value. The warranting of criteria in this study provides for a value factor to account for the demonstrated differences in importance of each criterion. This value factor for each criterion was established by assigning the weighted raw score totals of all groups for each criterion to that criterion. For ease in computation and handling, weighted scores have been identified as decimal value factors. Such value factors permit an evaluation instrument to be constructed including all or selected criteria from the study. An Effectiveness Evaluation Scale could use criteria from the research in the following fashion

Sample item:

Check the term which in your judgment best describes your professor's characteristic teaching behavior.

This professor demonstrates comprehensive knowledge of his subject.

Always___ Most of the time___ Occasionally___ Very Seldom___

Never____.

A student marking the space "always" would be giving the faculty member a "5" on that item which when multiplied by its value factor of .732 would give him a score of 3.66 on this one item.

The sum of the products of the criterion ratings and the criterion value factors would produce an effectiveness score.

Findings

1. All rank order correlations between selected groups of respondents are different from 0 at the .01 level of significance for individual criteria.

2. Sixty criterion behaviors associated with effective teaching at the University of Toledo have been established as warranted for evaluation of such teaching.

3. The academic community of the University of Toledo is agreed on the importance of the sixty criteria in the rank order which is established in the study.

4. A table of weights of importance has been established to account for the importance of each criterion.

5. Rank order correlations are different from 0 at .05 level of significance for the major behavior categories between 72 of the 78 selected groups.

Observations

Research on the effectiveness of teaching indicates promise in clarifying issues which surround this presently popular topic related to the evaluation of teaching. Such research can also help prevent the perpetuation of error in such evaluations or at least provide an analysis of a major problem in any evaluation. That problem is the identification of criteria to be used. This study seems to have done this for the present at the University of Toledo. An additional useful result of this study is the providing of a value weight for each criterion which could be used in an evaluation instrument in order that some accounting of the differences in importance of criteria used in such evaluations may be accomplished.

The study reported here is apparently unique in that it provides a sample of one institution's total academic community an opportunity to participate in consideration of criteria which may be used in evaluation of effective teaching. It is the only study apparently in which the judgments of a representative sample of a complex academic community on such criteria have been tested against a jury of outstanding teachers in an institution.

Of course, significant problems remain in the evaluation of effective teaching. They are:

1. The competence of persons doing the evaluation, and
2. The test of reliability of the criteria and procedures which can only be accomplished through longitudinal studies.

It seems though that a sound beginning has been established with the identification of criteria in this study.

Usefulness of the Findings of this Study

The University of Toledo was not completely satisfied with the fact that it had established, on statistical grounds, criteria useful in the evaluation of teaching and consequently we sought the assistance of three other universities who had indicated an interest in having the University of Toledo study replicated on their campuses. The criteria which had been established in the Toledo study were then placed in response instrument form and distributed to sample populations at the University of New Mexico at Las Cruces, Western Kentucky University, and Northern Illinois University. We did this because although the University had completed research which substantially identified criteria of effective teaching appropriate for the University of Toledo, the question remained as to how these criteria would fare under the evaluation of their warrantability in a wider form of judgment.

Invitations to participate in the research were sent to the Offices of Institutional Research at New Mexico State University, Northern Illinois University, and Western Kentucky University. These institutions agreed to participate in the research and accepted the offer of the University of Toledo to furnish the materials necessary for the research and the services required to process the data and interpret it. The same response instruments used at the University of Toledo in identifying the importance of each criterion behavior were prepared in quantities requested by New Mexico State, Northern Illinois, and Western Kentucky. These were given to the randomly selected sample populations at each institution in the Spring of 1968 with data being sent to Toledo for processing during the late Spring and over the Summer of 1968. The derived ranks for each criterion behavior by each university are shown in Table 1.

Insert Table I here

The four universities are in agreement that:

1. Each criterion behavior identified in the response instrument is warranted for the evaluation of effective teaching.

2. The criteria are important in the evaluation of such teaching in the rank order established by the study.

3. There is no significant disagreement among the reporting categories

selected for study about the rank order importance of these criteria.

The research effort over the past two and one-half years identified with this study has been fruitful particularly for the following reasons:

1. Apparently for the first time, large numbers of the significant segments of four universities have identified criteria warranted for the evaluation of effective teaching in their universities.

2. For the first time, four public universities have cooperated to test the findings of their individual research on effective teaching against the judgments of other academic communities.

3. The increasing acceptance of the results of this research by students and faculty is an indication that the procedures and findings are proving useful.

Those who have worked with the study for two and one-half years consider all of the above useful, satisfying, and one more small step toward the establishment of some better ground on which to evaluate teaching but by no means the end of such research. One cannot hope to establish a universal system for such evaluation. The possibilities provided in the procedures here indicate that since there is such high correlation in the judgments of these public universities that it can be hypothesized that similar results would be found in the responses from a larger number of public universities. If such were to be the case, we might be on the path to the identification of a typology of student and faculty who attend and teach at such institutions in terms of their attitudes toward effective teaching. Similar research conducted in the sector of private higher education or sectarian higher education might produce interesting and useful results.

The College of Education at the University of Toledo considered the results of the study soon after its completion in 1967, and, consequently changed their procedures of evaluation by incorporating the top 15 or 20 of the criteria in a newly designed evaluation instrument. The administration of that instrument and the research which has followed is described in a companion paper attached hereto.

STUDENT PERCEPTIONS OF TEACHING EFFECTIVENESS

Introduction

On the basis of the judgments of teaching effectiveness by several relevant populations, students, faculty, and alumni, as described in a companion paper, the College of Education, University of Toledo, prepared a 15-item rating scale. Fourteen of the items (later revised to nineteen) were chosen from those characteristics most often judged as critical in describing teaching effectiveness. The fifteenth item (later the twentieth) asked the student to provide a global rating of teaching by the instructor of the courses in which they were enrolled. It was expected that those items preceding the last item would provide a multi-dimensional frame of reference within which a mediated judgment of teaching could be obtained. (See Appendix for latest form used.)

The original intent of the scale was to provide a formal feedback routine for the instructors about their instructional methods. Both a summary of the ratings received from the students and their unstructured comments were given exclusively to each instructor. In the Fall of 1968, the College faculty voted to provide the information from the ratings to the elected College salary and promotions committee. Such decision brought about several problems. One of the major problems was that of preparing effective guidelines for the interpretation of numbers whose truth value did not extend to the fourth decimal place. Another one was that of reconciling the tendency of students to give poorer ratings if they were enrolled in large classes with other freshmen and sophomores, and to give higher ratings if they were enrolled in small classes with other graduate students.

In an attempt to diminish the bias present by size of class and instructional level, thirteen norm groups were established. The rating of each course then was made relative to the ratings of other courses of the same size and class level. That is, the rating of an instructor on "overall teaching" was transformed to a standard score using the appropriate mean and standard deviation. The average of these standard scores for each course for which the instructor was responsible became the index of "effectiveness" as perceived by the students.

The problem related to interpretation was answered by categorizing faculty indices into one of three classifications: upper one-fifth, middle three-fifths, and lower one-fifth. Such information was provided to the salary and promotions committee.

Construct Validity

As is undoubtedly well known, the study of the validity of a scale alleged to be measuring a construct is characterized by the relationships of the scores derived from the scale with other variables, variables with which the relationship is expected to be strong as well as variables with which the relationship is expected to be minimal or null. Several studies have been made with the Student Perception of Teaching Effectiveness Scale focusing on the latter set of variables--those with which the relationship is expected to be minimal. Essentially, the studies were those of bias. If the scale is valid, the relationship of the scores with grades, with class size, with instructional level, with sex, with G.P.A., and the like ought to be minimal or zero. The tables on the next few pages display the information collected.

Tables 2, 3, and 4 reveal information which suggests that factors besides "teaching effectiveness" are related to the scores derived from the Student Perception of Teaching Effectiveness Scale. Table 2 clearly indicates the bias extant in class size and instructional level. In a nearly perfect order, the rating increases in numerical size from smallest to largest size classes. Similarily, though not as perceptible, the general ratings by level of instruction increase in numerical size from graduate students to freshman-sophomore levels. The relationship between the interaction of these two variables and the scores derived from the scale has been measured as 0.11 (the correlation ratio--eta squared). Statistically one can remove the bias by "partialling" it out--by setting up separate norms.

Insert Table 2 here

Table 3 also reveals certain tendencies which would suggest a relationship between the variables and the scores from the scale. While the variable of sex of student and the required-elective variable seem to have but slight relationship, the "reported" GPA (reported = student reported) and expected grade indicate clearly discernible differences of mean ratings over the several levels of each. While the first two variables, sex and required-elective, can be diminished through norming procedures, the variables of GPA and grade are not so easily dismissed. The former is amenable to distortion by student manipulation and ignorance--consider the responses of 85 graduate students with respect to GPA who reportedly have received a pattern of grades which would clearly restrict them from attending classes. The latter variable is amenable to distortion

Insert Table 3 here

by act of the faculty member who assigns the grades. And grades seem to have considerable relationship with ratings. Table 4 indicates that

Insert Table 4 here

the average Pearson "*r*" within each norm grouping is *r* = -0.42. (Note that a negative correlation would indicate that the higher the grade received, the higher the rating given--negative because of the inverse nature of the meaning of the scale orders for the two variables.) A bit of explanation is in order about the technique employed to obtain the correlations shown. The elements in the calculation are class characteristics, not individual student characteristics. Each class or course received an average student rating; each class also was categorized by the average grade given by the instructor to the students enrolled therein. Then, within each norm group, and later within each instructional level, the Pearson "*r*" was obtained.

The size of the correlations is quite striking. To be sure, what is offered is a record of but one administration of the scale--Spring, 1972. Yet, correlations of -0.78, -0.77, -0.60, -0.59, and -0.55 are so large that it would be quite unexpected for them to vanish in another administration of the scale. The indictment of the validity is very strong; what the correlations reveal is that the variations in course ratings is accounted for to the extent of from 30 to 60% by the grades assigned. One could argue that those who give higher grades are those who are more effective; yet, it would be difficult to convince those who reportedly have the same students in their courses and have a different line on grades. Whatever, this problem must be resolved in some fashion before one can build a reasonable case for validity.

Scale reliability. The question of reliability of the outcome of the scale administration has been given but cursory examination.

The question of reliability is not that of the usual "individual" assessment but that of the average assessment. It would appear that where there is considerable consensus on the rating to be given, to that degree there is some confidence in the reliability of the average obtained. Where there is a lack of consensus e.g., a uniform distribution of ratings, less confidence appears warranted. A study of the extreme fifths and middle three-fifths of the distribution of standard scores, referred to earlier as indices of effectiveness, revealed that the order of consensus is directly related to the order of "effectiveness." The median modal relative frequency for the upper one-fifth was 86%; for the middle three-fifths, 58%; for the lower one-fifth, 42%.

Other studies. Other studies have had little central focus but to pursue "interesting" questions. A factor analysis of the scale was undertaken to note (1) whether we were measuring a unitary trait, and (2) if not, what factors appeared to be present in the set of items. The following clustering of items or topics were determined:

Knowledge and Skill in Explanation

Meaningful class preparation
Interest in subject
Knowledge of subject
Motivation of students
Ability to explain
Responses to questions
Overall teaching

Concern for Students

Fairness in evaluation
Respect for students
Availability for consultation
Promptness in returning assignments
Offer of assistance

Use of Teaching Tools

Examinations required understanding
Fairness in evaluation
Value of textbook
Overall teaching

Inspiration

Encouraged independent thought
Motivated students
Respect for students

Pressure to apply the means of evaluation often stem from some dissatisfaction of that which is to be evaluated. That is, evaluation should in some way improve the quantity or quality of the item or process.

Table 5 portrays the experience of the College with average student ratings for the fifteen item (later twenty item) scale since the Spring of 1968.

Insert Table 5 here

(The fifteenth question and the twentieth question from the initial and revised scales, respectively, are identical--thus the peculiar format used in the last three columns of Table 5.) It is noteworthy that the perception of "overall teaching" and other items have tended to improve, albeit, somewhat irregularly. To the degree that student's perceptions are accurate, the evaluation routine has had a beneficial effect.

Summary. The College faculty, as a group, has recently confirmed their opinion that the information obtained through the use of the Student Perception of Teaching Effectiveness Scale is useful in deliberations of the Salary and Promotions Committee. That is, such information has greater validity than the "gossip" which formed the basis previously for such deliberations. It is likely however, that those who make the decisions are not cognizant of the caution necessary in the interpretation of the information given. We are hopeful that studies that we can generate together with the information available from others can improve our confidence in our results and the decisions made. It would be extremely useful to have access to a "clearing-house" which allowed the concentration of information and the dissemination required to make progress.

References

Biddle, Bruce J. *Contemporary research on teacher effectiveness.* New York: Holt, Rinehart, and Winston, 1965.

Fattu, N. A. Research on teacher evaluation. *The National Elementary Principal*, 1963, *63*, 19.

Gage, N. L. Address appearing in *Proceedings*, Research Resume 1960. California Teachers Association.

Holland, John L., and Richards, James M. ACT Research Report No. 8, 1965.

McKeachie, W. J. In N. L. Gage (Ed.), Handbook of research in teaching.
Chicago: Rand McNally, 1963.

Table 1

Rankings of Criterion Behaviors by Institution

	Criterion Behavior	Rankings			
		NMS	NTU	WKU	UT
1.	Evidencing better than average speech qualities	26	25	27	26
2.	Constructing tests which search for understanding on the part of the students rather than rote memory ability	4	5	9	5
3.	Providing several test opportunities for students	27	28	29	32
4.	Engaging in continued formal study in his field	24	29	31	28
5.	Acknowledging all questions to the best of his ability	12	12	14	12
6.	Motivating students to do their best	11	9	5	10
7.	Explaining grading standards	40	37	42	45
8.	Publishing material related to his subject field	57	59	60	60
9.	Having practical experience in his field	20	19	24	21
10.	Communicating effectively at level appropriate to the preparedness of students	7	6	6	7
11.	Identifying his comments which are personal opinion	28	27	41	27
12.	Challenging students' convictions	44	38	52	43
13.	Utilizing visual aids to assist in creating subject matter achievement with students	47	48	45	47
14.	Announcing tests and quizzes in advance	39	41	36	46
15.	Making written comments on corrected returned assignments	22.5	17	26	25
16.	Presenting organized supplementary course material	43	42	47	41
17.	Establishing good rapport with students in classroom	17	15	15	17
18.	Making an effort to know students as individuals	36	30	28	38
19.	Inspiring students to continue for graduate study	52	52	51	49
20.	Demonstrating comprehensive knowledge of his subject	6	10	10	3
21.	Exhibiting an intelligent personal philosophy of life	46	44	38	40
22.	Encouraging student participation in class	25	22	23	24
23.	Beginning and ending classes on time	48	51	48	51
24.	Accepting justified constructive criticism by qualified persons	22.5	21	21	23
25.	Sharing departmental duties with his colleagues	50	49	49	50
26.	Having irritating personal mannerisms	53	54	57	54
27.	Establishing sincere interest in subject being taught	2	2	3	2
28.	Taking measures to prevent cheating by students	38	43	32	31
29.	Recognizing his responsibility for the academic success of students	21	26	17	18
30.	Devoting time to student activities on campus	59	58	54	58
31.	Demonstrating a stable level-headed personality	35	31	22	30
32.	Returning graded assignments promptly	30	32	34	34
33.	Patiently assisting students with their problems	16	18	13	20
34.	Holding membership in scholarly organizations	55	55	56	56

Table 1 (Continued)

			Rankings		
	Criterion Behavior	NMS	NIU	WKU	UT
35.	Being well prepared for class	1	1	1	1
36.	Setting high standards of achievement for students	18	23	25	16
37.	Involving himself in appropriate university committees	58	56	55	55
38.	Being knowledgeable about the community in which he lives	54	53	53	53
39.	Being readily available for consultation with students	14	14	16	15
40.	Displaying broad intellectual interests	41	36	40	36
41.	Treating students with respect	10	4	2	11
42.	Raising the aspirational level of students	19	20	18	17
43.	Being able to show practical applications of subject	13	13	12	13
44.	Organizing the course in logical fashion	8	11	11	9
45.	Making appearances which assist programs of community	60	60	58	59
46.	Earning the respect of his colleagues	45	45	37	42
47.	Encouraging intelligent independent thought by students	5	7	8	8
48.	Using teaching methods which enable students to achieve objectives	9	8	7	4
49.	Rewriting and updating tests	15	16	19	14
50.	Presenting an extensive lucid syllabus of the course	49	46	50	48
51.	Explaining grading procedures	37	34	39	41
52.	Being consistently involved in research projects	56	57	59	57
53.	Seldom using sarcasm with students	34	47	43½	39
54.	Indicating that the scope and demands of each assignment have been considered carefully	33	35	35	33
55.	Being fair and reasonable in evaluation procedures	3	3	4	6
56.	Relating course material to that of other courses	31	40	46	35
57.	Using more than one type of evaluation device	29	24	30	29
58.	Being neatly dressed	51	50	43½	52
59.	Exhibiting a genuine sense of humor	42	33	33	37
60.	Encouraging moral responsibility in students by example	32	39	20	22

Note - NMS = New Mexico State, Las Cruces; NIU = Northern Illinois University; WKU = Western Kentucky University; UT = University of Toledo. The number of responses on which the above information is based: NMS = 654; NIU = 2488; WKU = 1698; UT = 1793.

Table 2

Mean Ratings and Standard Deviations by Size of Class and Instructional Level

Instructional Level	Size of Class				
	1 - 10	11 - 24	25 - 49	50 - 100	OVER 100
Graduate	M = 1.33 SD = 0.50	M = 1.55 SD = 0.41	M = 1.69 SD = 0.48	M = 1.65 SD = 0.63	
Junior - Senior	M = 1.63 SD = 0.48	M = 1.76 SD = 0.49	M = 1.76 SD = 0.57	M = 2.24 SD = 0.41	
Freshman - Sophomore	M = 1.48 SD = 0.52	M = 1.67 SD = 0.40	M = 1.83 SD = 0.50	M = 1.85 SD = 0.52	M = 1.91 SD = 0.35

Note - Ratings are based on a scale of 1 - 4, 1 is labeled excellent, 4, poor. Means and standard deviations shown have been accumulated to Spring, 1972.

Table 3

Mean Ratings for Student Perceptions of Teaching Effectiveness Given Certain Characteristics of Class Members.

Characteristic	Freshman - Sophomores N	Freshman - Sophomores Rating	Juniors - Seniors N	Juniors - Seniors Rating	Graduates N	Graduates Rating
Required Course	718	1.731	549	1.607	437	1.556
Elective Course	196	1.714	143	1.531	249	1.558
Males	369	1.751	247	1.664	321	1.517
Females	562	1.740	451	1.567	376	1.614
Reported G.P.A.						
0.00 - 2.00	139	1.604	44	1.545	34	1.529
2.01 - 2.50	283	1.756	160	1.581	51	1.745
2.51 - 3.00	217	1.806	219	1.543	55	1.509
3.01 - 3.50	197	1.802	199	1.643	200	1.570
3.51 - 4.00	61	1.836	55	1.764	315	1.549
Expected Grade						
A	356	1.632	369	1.466	433	1.513
B	333	1.775	205	1.693	158	1.677
C	149	2.007	65	2.108	19	1.632
D	26	2.308	18	1.889	22	1.864
E	11	1.818	11	1.364	22	1.636

Note - N is the number of individuals in such classification who made a rating in the Spring, 1972. The base for the rating is: 1 - Excellent, 2 - Good, 3 - Fair, 4 - Poor.

Table 4

Correlations of Mean Grades Given in Course and Average Ratings on Overall Teaching by Norm Grouping

Spring, 1972

Instructional Level	Size of Class				
	1 - 10	11 - 24	25 - 49	50 - 100	OVER 100
Graduate	N = 20 r = -.03	N = 18 r = .07	N = 14 r = -.28	N = 0	N = 0
	Graduate, combined r = -.08 Graduate, size partialled out, r = -0.08				
Junior-Senior	N = 10 r = -.78	N = 17 r = -.60	N = 16 r = -.59	N = 0	N = 0
	Junior-Senior, combined r = -0.63 Junior-Senior, size partialled out, r = -0.65				
Freshman-Soph.	N = 7 r = -.77	N = 13 r = -.26	N = 8 r = -.55	N = 0	N = 0
	Freshman-Sophomore, combined r = -0.32 Freshman-Sophomore, size partialled out r = -0.40				
Total Group	Combined r = -0.42 With size and level partialled out, r = -0.42				

Note - scales measuring rating of class and grades are inverse in meaning-- "1" is the best score on rating scale, "4" the poorest; "4" is the highest score for G.P.A., "1" is a lower score.

Summaries of Rating for Spring Quarter, 1972, and Experiences from Other Years

Item No.	Spring Qtr. 1972	Fall Qtr. 1971	Spring Qtr. 1971	Fall Qtr. 1970	Spring Qtr. 1970	Fall Qtr. 1969	Fall Qtr. 1968	Spring Sem. 1968
1	1.6795	1.6773	1.7440	1.7942	1.7323	1.7850	1.7770	1.6494
2	1.3150	1.3108	1.3599	1.4284	1.3695	1.3885	1.3638	1.4155
3	1.3637	1.3893	1.4382	1.4953	1.3839	1.4611	1.4270	1.7516
4	1.8484	1.9629	2.0268	1.9223	2.0247	2.0880	2.0615	2.0347
5	1.6987	1.7548	1.8534	1.3333	1.8278	1.8172	1.7587	1.9001
6	1.6999	1.7622	1.7976	1.8348	1.7607	1.8206	1.7769	2.0763
7	1.8801	1.9255	1.9862	2.1008	2.0577	2.0906	2.0756	1.9554
8	1.5057	1.5037	1.5816	1.6355	1.5803	1.5774	1.5297	1.7624
9	1.5925	1.6299	1.7196	1.7586	1.6791	1.7215	1.7153	1.6985
10	1.4170	1.4401	1.5312	1.5768	1.4851	1.5118	1.4832	1.6431
11	1.7016	1.7225	1.7517	1.7927	1.8514	1.8509	1.7068	1.8460
12	1.6628	1.7704	1.7674	1.7583	1.8941	1.8418	1.7771	1.8631
13	1.5720	1.6430	1.6781	1.7450	1.7534	1.8250	1.6705	2.0948
14	2.0619	2.1683	2.2065	2.2249	2.2440	2.3103	2.3383	2.1984
15	1.7150	1.7483	1.8285	1.8920	1.8020			
16	1.6193	1.6914	1.7458	1.7981	1.7512			
17	1.8468	1.8596	2.0000	1.9750	1.9835			
18	1.7639	1.7781	1.9318	1.9761	1.8969			
19	1.7302	1.7652	1.8882	1.9540	1.8802			
20	1.6460	1.6661	1.7825	1.8250	1.7432	1.8131	1.7853	1.8236

BEST COPY AVAILABLE

Appendix

IBM 807 PRINTED IN U.S.A.

DIRECTIONS: Listen carefully to the instructions offered by the administrator of this rating scale. When you have decided upon the proper response, blacken the appropriate spaces on this sheet with a PENCIL. Make your mark as long as the pair of lines, and completely fill the area between the pair of lines. If you change you mind, erase your first mark COMPLETELY. Please make no stray marks for they may be misinterpreted. (More information is offered below)

IDENTIFICATION NUMBER

A	0	1	2	3	4	5	6	7	8	9
B	0	1	2	3	4	5	6	7	8	9
C	0	1	2	3	4	5	6	7	8	9
D	0	1	2	3	4	5	6	7	8	9
E	0	1	2	3	4	5	6	7	8	9
F	0	1	2	3	4	5	6	7	8	9
G	0	1	2	3	4	5	6	7	8	9
H	0	1	2	3	4	5	6	7	8	9
I	0	1	2	3	4	5	6	7	8	9
	0	1	2	3	4	5	6	7	8	9

1. Meaningful class preparation and planning. . . . .
2. Demonstrated sincere interest in the subject . . . .
3. Demonstrated comprehensive knowledge of the subject. . . . .
4. Employed exams which required understanding of ideas in course . . .
5. Demonstrated fairness and reasonableness in evaluating students. . .
6. Encouraged independent thought by students . . . .
7. Motivated students to do their best . . . .
8. Demonstrated respect for students . . . .
9. Demonstrated ability to explain course material. . . . .
10. Responded to questions to the best of his ability. . . . .
11. Availability for consultation . . . .
12. Demonstrated promptness in returning graded assignments and exams.
13. Offered assistance to students with problems connected with course .
14. Value of text and/or instructional materials used. . . . .
15. Personal interest and sensitivity to student problems. . . . . . .
16. Assignments made were in harmony with course objectives. . . . . .
17. Outside work demanded in line with course credit hours . . . . . .
18. Rating of course content as unportant and valuable . . . .
19. Class experiences have promoted learning on your part. . . . . . .
20. Overall evaluation of teaching in this course. . . . .
21. Course: Required Col. 1, Elective Col. 2. . . . .
22. Sex: Male Col. 1, Female Col. 2. . . . .
23. G.P.A. (See Below). . . . .
24. Expected Grade (See Below) . . . .

Rows A, B, and C. Code Number of Instructor.
Rows D, E, and F. Course Number
Rows 1, 2, 3, ...19, 20. Excellent, First Column - Good, Second Column
Fair, Third Column - Poor, Fourth Column

G.P.A.

0.00 - 2.00	1st column	A.......1st column	
2.10 - 2.50	2nd "	B.......2nd	"
2.51 - 2.50	3rd "	C.......3rd	"
3.01 - 3.50	4th "	D.......4th	"
3.51 - 4.00	5th "	F.......5th	"

Column Five (5) Should Not Be Marked. If Not Applicable Leave Blank.

1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40

CORRELATES OF STUDENT RATINGS

W.J. McKeachie

University of Michigan

What factors influence or are correlated with student ratings of teachers? Most of the work on student ratings has been strictly empirical--beginning with general notions about what a good teacher ought to do, writing items about these characteristics, factors analyzing them, and then attempting to validate them. But to understand what these ratings mean we need to fit them into larger theoretical structures. One way of doing this is to relate them to other variables that we know something about.

Basically we assume that student ratings are descriptive of teacher behavior and of the teacher's effect upon the student who fills out the rating scale. Insofar as the items of the scale are descriptive of teacher behavior we expect high inter-rater agreement, but we expect greater valid (and invalid) variability when we ask for the students' assessment of teaching effectiveness or value of the course in their own education. Some of us are following Dr. Hoyt in trying to get a clearer picture of what goes into such an overall rating by asking about the effect of the course on the student's judgment of his achievement of several different kinds of goals. From all that is known about social perception and attitudes, it seems very unlikely that judgments of teaching effectiveness are unaffected by student characteristics. Thus it is important to know what student characteristics affect ratings and the degree to which a given set of ratings are the result of autochthonous factors rather than of the more objective qualities the rating was intended to assess.

Another set of characteristics likely to influence student ratings of teaching are characteristics of the class or course. Is it easier to get good ratings in small classes than in large? Does the teacher of a required course have a tougher job than the teacher of an elective course? In interpreting a teacher's ratings we usually are influenced by some assumptions about such variables. Thus it is important to know how valid these assumptions are.

A third set of factors influencing ratings are characteristics of the instructor himself. Does an experienced teacher get better ratings than an inexperienced one? What personality characteristics of the teacher influence what he does in teaching and how the students react to him?

In this paper I do not intend to review interrelationships between items on scales for student ratings of teaching nor will I enter the realm of correlations between student ratings and student learning or other criteria of validity. Each of these topics would constitute a paper in itself.

Student Characteristics

The classic research on most aspects of student ratings of instruction was carried out by Herman Remmers and his students at Purdue. His results are still largely unchallenged by more recent research. Among the factors which did not significantly affect student ratings were such student characteristics as:

Veteran/non-veteran status

Age

Sex

Class standing

Grade in course (However when the top students achieve more than expected they rate the course higher, and when the poorer students do better than expected they rate the course higher).

Student Characteristics

S expectations Kelley and Perry, Niemi, & Jones (1973) have shown that student expectations affect ratings for a single lecture, but we have little evidence on the dynamics affecting persistence of expectancies over a term.

Personality Costin & Grush (1973, unpublished) found no relation between traits measured by the Gordon Personality Profile and ratings. The organizers of this conference hoped to stimulate research, so we did some. Our results, like those of Costin & Grush, were largely negative. Using Gough's California Psychological Inventory (the CPI) we obtained only three significant interactions out of fifty tested.

Content

Carney and I found some interaction between content and sex affecting ratings in a psychology course. Women like life-oriented topics; men liked science-oriented topics. Turner *et al* found that high anxiety students prefer personality-social content.

Course Characteristics

Class size Generally smaller classes are preferred, but results are not uniform. Often the best teachers are assigned larger classes and are rated well. Perlman (1973) found that students at Manitoba rated smaller classes higher on two major dimensions--intellectual stimulation and socio-emotional climate.

Required vs. elective   Remmers found no difference, but Lovell & Hane. (1955) and Kapel here at Temple found that required courses were rated lower.

The relatively small effect of variables such as size or required vs. elective lead me to feel that students take account of the teacher's task in their ratings. They may give higher ratings if they think a course is hard to teach. Moreover they may give higher ratings if they can assess their learning in conventional ways. Hence, there may be a bias toward lecture-test courses which is not reflected in real long-term effects. Shillace, for example, reports 95% retention of anecdotes; 25% retention of the point of the anecdote in lecture.

Students can judge whether they followed a lecture and can count pages of notes.

Students are less likely to be able to evaluate gains in ability to analyze or evaluate. The fact that difficulty of a course has no effect on ratings is not as surprising as it may seem. There are many ways of making a course difficult, most of which have little to do with increased learning. Moreover, students despise Mickey Mouse courses. As Walster has shown in laboratory experiments,"hard-to-get" goals are rated higher. Students may neglect to include in their rating skillful planning of method, content, textbook, teaching technology. But they do give credit for trying, for concern.

Instructor Characteristics

Sex - No difference (cf. Centra)

Age - Younger teachers are rated higher (Riley, 1959)

Rank - Results are mixed

Degree - BA instructors are rated lower than MA's or PhD's (Riley, 1949)

Experience - Mixed results, but mostly some improvement with experience (Costin)

No effect (Centra)

Grading standards - Mixed results on this in terms of overall ratings, but lower graders are rated lower on fairness of grading (Heilman & Armertrout, 1966)

Knowledge of subject - No effect

Knowledge of teaching - No effect

Research - Publishers not higher (Aleamoni). Second authors are rated higher. First authors of books were rated poorly (Feldhusen)

Personality of Instructor

Getzels & Jackson (1963) reviewed 150 studies (public schools) and concluded that little is known about instructor personality.

The same is true of instructor personality and ratings at college level. Bendig (1955) and Sorey (1968) found no relationship between Guilford-Zimmerman scores and effectiveness.

In our studies at Michigan peer ratings of the general culture of a teacher correlated positively with student learning and ratings. Enthusiasm-Surgency on the 16PF was also positive.

Costin & Grush (1973) using the Gordon Personal Profile found that vigor and student-perceived original thinking, personal relations, and ascendency were also positively correlated with student-rated effectiveness.

Discussion

The results of these studies contain both good news and bad news. A lot of variables that might affect ratings don't have much effect. This is good news in that a number of potential sources of error are thus determined to be of little consequence and those of us reporting ratings to instructors need not worry about constructing different sets of norms for particular kinds of classes or particular kinds of students. The bad news is that my hope that these correlates would lead to new theoretical insights is also not supported. Intuitively, one feels that one needs to separate the effect of the teacher as he teaches in the classroom from that of the teacher as a person. Each of these must make some impact upon the student and in turn upon his ratings. They must have some differential effect on different types of students. My faith in the usefulness of such detailed analysis remains despite, not because of the richness of, our research findings to date.

I still believe that teaching is a very complex business. Thus I think interpretation of student ratings should be left to faculty who understand the particular problems of a particular class and who can make allowances for the variables which might affect student judgments.

Peers may over-weight some factors, hence our research is worthwhile to them. But teaching is still a very human and individual endeavor and its meaning is not easilty captured by statistics.

TEACHERS WHO MAKE A DIFFERENCE

Jerry G. Gaff

California State College, Sonoma

Two basic ideas which underlie the use of student ratings are that systematic procedures should be used to evaluate teaching effectiveness and that students should play an important part in that process. These twin assumptions have been operationalized in the form of student ratings of their teachers, and the solicitation of such ratings is not at all uncommon these days.

However, most teaching evaluation procedures are quite modest efforts. Most (a) rely on student descriptions of their teachers, (b) in a classroom, (c) for the duration of a term, (d) at the discretion of individual faculty members. This despite the fact that it is obvious that (a) students are but one constituency with a legitimate interest in and perspective on the quality of teaching, (b) the classroom is only one setting in which teaching and learning occur, one which may be becoming decreasingly important, (c) the important consequences of an education can be observed only over a long time span, and (d) acquisition of knowledge about the effects of one's teaching can help all teachers learn how to improve.

The main thrust of my comments today is that it is necessary to go beyond this current limited use of student ratings. I am prepared to argue that it is important to advance in three areas -- in research, in theory, and in practice.

First in regard to research. It must be acknowledged that even the modest initial efforts to evaluate teaching have generated several useful student rating forms, many research studies, and a number of correlates of effective teaching. Despite these advances, however, very little is known about the characteristics of teachers, teaching styles, and student-teacher relationships which have demonstrable long term benefits to students. We need to conduct research which will provide knowledge about the kinds of teachers and teaching which make a difference in the cognitive and affective lives of students. This kind of research probably will have to employ methodologies beyond those which are commonplace in the study of student ratings. I would like to illustrate the kind of research which is needed by discussing a study which I have recently completed.

While working at the Center for Research and Development in Higher Education at the University of California, Berkeley, I was presented with a special opportunity to examine the impacts of faculty on students during their entire four year career. Longitudinal studies of student growth and development were initiated in 1966 under the general direction of Paul Heist. These researchers administered a set of questionnaires to students when they entered as freshmen and again as they were preparing to graduate in the spring of 1970. In conjunction with these studies, several colleagues and I, who had been researching faculty members, conducted a survey of faculty in nine of the same institutions during the spring of 1970. We related data from 851 faculty members to 1475 students for whom complete sets of freshman and senior questionnaires were available.

Of particular concern to all of us were certain kinds of teaching and learning, those which are usually lumped together under the term "liberal education." Although that term cannot be defined sharply, it manages to imply a special kind of education which is at the heart of most college and university endeavors. In regard to teaching, it means more than transmitting facts and theories and more than presenting the content of one's academic specialty, however important these may be. It implies a breadth of concern and an attempt to relate knowledge in one's field to other fields of investigations, to realities in the larger society, and to the personal lives of students. Similarly, the kind of learning which is the fruit of a liberal education transcends the acquisition of cognitive facts, methods, or principles, as important as these may be. It includes such affective components as acquiring an appreciation of the value of intellectual inquiry, increasing sensitivity and awareness, and developing a personal philosophy and outlook on life. In short, the kind of teaching and learning in which we were interested was that which made a difference in the lives of students.

From the mass of data which were gathered several analyses were conducted, but I will discuss only a couple today. One item asked senior students to name the faculty member who had "contributed the most to their educational and/or personal development" during their college years and to describe the ways thatthe teacher had helped them A total of 1127, 77 percent, of the seniors named such faculty members. Most of the remaining students in the survey left the item blank, but a few wrote in colorful comments like "No such animal," disavowing that any faculty member had played a significant role in their development.

Insert Table 1 about here

As may be seen from Table 1, the vast majority of the nominated faculty were said to have been available and open for discussions, stimulated students intellectually, helped them feel confident, demanded high quality work, and interested students in their fields. Fewer, but still a majority of the influential teachers, were said to have encouraged students to inspect their values, given career advice, and fostered awareness of social issues. Only a minority counseled about a personal problem or helped students get a job or scholarship. Although a couple of these statements are descriptive of the teachers, most are descriptive of the ways students were helped by them. Generally, the results confirm that students benefited in several ways and to a considerable extent by the teachers named.

However, this is only the prologue to the issue at hand, because we wanted to learn about the kinds of teachers who had such impacts on students. Seniors were also asked to name, but not to describe, the teacher who had taught the most "stimulating course" they had taken during their college careers. A total of 97 faculty members who received nominations from two or more students either as having contributed the most or as the teacher of the most stimulating course had returned faculty questionnaires. A total of 609 faculty who received no nominations in either capacity also returned questionnaires. In one analysis the responses to the faculty questionnaire of these two groups were contrasted.

Similarly, faculty members were asked to name two colleagues whom they regarded to be "outstanding teachers" and one colleague whom they regarded as having "significant impact on the lives of students." Another analysis contrasted the questionnaire responses of the 137 faculty members who received two or more nominations from their colleagues with the 525 who received no mention from any colleague.

We first discovered that there was a fair degree of overlap between the faculty nominated by students and those named by colleagues. This overlap helps to explain why the results of the two analyses are so similar that they can best be discussed together.

More importantly, we learned that there is a configuration of variables which differentiates the faculty who make a difference from the rest of their colleagues. First, faculty nominated *both* by students and by colleagues evidenced a greater commitment to undergraduate teaching than did the non-nominated groups. In significantly greater numbers they registered preferences for teaching over engaging in research and for teaching undergraduate students over graduate students.

Influential teachers were also significantly more likely to talk with students about a variety of issues of importance and even urgency to young adults. In both the student- and colleague-nominated analyses, over 50 percent of the influential faculty scored in the top third of a scale concerning the frequency with which they discuss with students youth culture issues, such as sex and morality, the use of drugs, and alternative life styles, whereas less than a third of the non-nominated faculty reported frequent discussions of this type with students. Such "rap sessions" -- whether they occurred inside or outside the classroom -- are evidence of the influential faculty's greater involvement with students and their greater concern for issues of importance to students.

In order to sharpen the interpretation of this finding, it should be noted that the nominated faculty were *not* more liberal than their less influential colleagues. A variety of issues which range along a liberal-conservative dimension including political preference, views concerning the regulation of student social life, tolerance for controversial activities of students and faculty, and attitudes toward student participation in policy-making failed to differentiate the two groups. Further, the student-nominated group of teachers did not differ in age from the non-nominees; influential teaching was found by students to be about equally distributed throughout the age groups. Thus, it was not the radical young faculty who were regarded as influential by discussing these youth culture issues with students, as might have been suspected. Rather, it appears that a willingness on the part of a teacher to explore and analyze these topics with students regardless of his age or position regarding them is the key to being regarded a particularly influential teacher by one's students and colleagues.

The single biggest difference between influential faculty and their colleagues was the extent to which they interacted with students outside the classroom. Faculty respondents were asked to indicate how many times they had out-of-class discussions with students in several areas ranging from course work to personal problems. Fifty-four percent of the student-nominated faculty scored high on the scale of frequency of such interaction compared with 30 percent who received no nominations; comparable figures for the colleague-nominated group were 55 and 26 percent. Perhaps encounters which take place outside of class provide greater opportunities for students and teachers to carry on discussions which focus on student concerns than the more formal student-faculty relationships which are found in the classroom. At any rate, these data indicate that much effective teaching can be found in settings beyond the classroom.

If making a difference with students can be thought of as constituting its own reward, then influential teachers would appear to reap a greater sense of accomplishment from their teaching efforts. Forty-four percent of the student-nominated faculty scored high on a scale of self-perceived influence which measured the extent to which faculty thought they had a impact on students' personal philosophies, decisions about careers and major fields of specialization, and appreciation of the values and methods of scholarly inquiry; only 27 percent of the non-nominated faculty felt they had as much influence on students generally. Comparable differences were found between those faculty receiving two or more nominations from their colleagues and those receiving none. Similarly, over two-thirds of each group of the influential teachers named a senior to whose educational or personal development they felt they had contributed a great deal, which was considerably more than the non-nominated groups. Given that non-nominated faculty had much less contact with students outside of class, it may be that they often did not know their students well enough to assess their own impact on them.

One finding that is particularly relevant to the concerns of this conference is that the nominated faculty generally were not distinguishable from their non-nominated colleagues on the basis of their classroom teaching styles. Thirty-two items descriptive of classroom teaching styles were included in the faculty questionnaire. Most of them were taken from the well developed and validated student rating scale developed by Hildebrand and Wilson and were modified so that faculty could describe their own teaching behavior. Reliable scales were developed to measure the extent to which faculty encouraged students to participate in the course, classes were well organized, teachers adopted a relaxed, discursive style, and faculty attempted to make their presentations interesting. Only the latter scale yielded statistically significant differences between nominated and non-nominated groups, and those differences were so small as to be educationally insignificant.

Here then is an interesting anomaly. Hildebrand and Wilson have developed one of the best student rating scales around; they have conducted research which demonstrates that its five scales consistently discriminate between effective and ineffective classroom teachers; but items borrowed from that instrument failed to differentiate between the teaching practices of faculty who make the greatest difference in the lives of students and their less influential colleagues.

How may this finding be explained? Of course, it may well be that the items were changed in meaning when they were modified for use with the faculty, and it is clear that the scales derived from the faculty data are not directly comparable to those derived from student data. But I am bothered by another possibility, the different contexts of the studies. Research into student ratings is generally conducted within the framework of a single course. So far as learning the subject matter of a course is concerned, the degree of organization, for example, may be a significant teaching factor. However, so far as making a difference in the lives of students is concerned, the degree of teacher organization would be trivial. Although it is by no means conclusive, this finding suggests to me that what goes on within individual classrooms may have little relevance for the long-term liberal education of students. If this is so, a research procedure designed to identify the correlates of effective teaching within the context of conventional academic courses may systematically fail to identify the kind of teaching which facilitates a liberating education.

There are many other analyses which I would like to share with you, but since time is lacking, you might find the essence of the study useful. The general conclusion of our study is that on most campuses there are important barriers to significant encounters between students and teachers. Those teachers are most influential who find ways to transcend the barriers of age and authority, classroom and content, to confront students where they are. Although I have not discussed it today, those students are more effective in reaping the benefits of a liberal education who more aggressively use the learning resources of the school, including their teachers, to expand their understanding and awareness. And those schools are most potent which, whatever the content of their formal curricula, create the conditions for casual, frequent, continuous, and wide ranging interactions between students and teachers which extend beyond the classroom.

I hope this brief description of a portion of one research effort will illustrate my major point that more research needs to be directed at the kinds of teaching which are associated with long-term beneficial effects on students. There are many kinds of teaching and many kinds of learning, and there is a need to learn about the qualities of teachers who make a difference in the cognitive and affective aspects of students.

You will recall that I said we need to go beyond the current state of the art of student ratings in research, in theory, and in practice. If the discussion about research was rather lengthy, the issue about theory may be handled with dispatch. The simple fact is that we lack an adequate theory of instruction. Research has identified various kinds of effective teaching and, as we have heard, several of its correlates. But we are not at all sure how the instructional behavior of teachers relates to learning by students. Given this lack of understanding, it is uncertain how the behavior of teachers may be modified to increase the amount of student learning.

The lack of adequate theorizing is particularly apparent in the area of student ratings. So far as I know, there is no theory which relates student ratings to instructor behavior, to changes in instructor behavior, or to student cognitive and affective growth. These theoretical questions must be addressed:

1. How are the results of student ratings conceived and interpreted by teachers? Do faculty selectively perceive the results, and if so, what needs does that process serve?

2. How do faculty members respond to the positive and negative results of students ratings? How do ratings affect their self concepts as persons and as teachers?

3. How do ratings affect faculty motivations to change their teaching behavior? Do negative ratings generate anxiety or other defensive reactions which impede change, or do they generate a genuine desire to improve the quality of teaching?

4. How do changes in teaching behavior affect students? Do the students perceive changes in their teachers, how do they respond to those changes, and do they learn more?

Unless we are able to improve our theorizing about the role of student ratings in the teaching-learning process, I see little hope that we can use them to help teachers make a greater difference.

The third area I want to comment on is current practice concerning student ratings. Even though we lack the desirable knowledge base in research and theory, the state of the art of student ratings is sufficiently advanced that we may go beyond the usual current practice. After all, most decisions we face in life must be made with only incomplete knowledge, and on the basis of my own impressions I will suggest a few guidelines for implementing a more comprehensive teaching evaluation procedure.

1. A formal system of teaching evaluation should be established for all faculty members simply because it may provide the best available knowledge about the consequences of teaching, and because such knowledge is necessary if faculty members are to improve their performance. In

order to make the evaluation the integral part of the instructional process it deserves to be, it may be possible to place the responsibility for obtaining reliable knowledge about his teaching on each faculty member.

2. Although student ratings are better than less systematic attempts to learn of student reactions to teachers, and although the evidence indicates considerable overlap between student and faculty judgments, the student viewpoint is only one which needs to be considered. A more comprehensive teaching evaluation procedure which solicits appropriate inputs from students, the teacher himself, his teaching colleagues, and administrators -- all of whom have legitimate interests in the quality of instruction -- would seem to be a more desirable procedure.

3. Evidence about the classroom performance of teachers is important but not sufficient, as the research data I have discussed indicates. It is particularly important to learn about how faculty interact with students beyond the classroom. Indeed, recent years have seen the advent of a number of new settings for teaching --independent study, community action projects, work-study programs, experiential learning, external degree programs -- in which the traditional classroom plays a more limited role. Evidence about the kinds of teaching which occurs in these expanded contexts must also be taken into consideration in teaching evaluation procedures.

4. Rather than a one-shot affair, teaching evaluations should be conducted on a continuous basis. A regular and continuous procedure would identify the degree of progress, stability, or even regression in performance and point the way for various actions which might assist each person to achieve to his fullest.

5. Although it is useful for faculty members to learn of the results of their own evaluations, it is more useful for them to learn about their own evaluation in comparison with the evaluations of others.

Insofar as possible, teaching evaluation should be conducted on a comparative basis.

6. Some individuals may make incorrect interpretations of their assessments, because they are not sophisticated in reading such data or their feelings may interfere with their understanding. For these reasons it may be desirable to build in follow-up procedures in which the results of teaching evaluation may be discussed, interpreted, and implications for changes (if any) drawn with the teacher. Such counseling would obviously be a delicate matter, but it can be used to assist teachers make good use of the assessment data.

7. One of the stickiest issues concerning evaluation concerns the use to which the results are put. It seems to me that the most important use is for them to be linked together with a faculty development program. A full-fledged faculty development program would be designed to assist individual faculty members to develop to their fullest both professionally and personally. There ought to be a variety of resources available at an institution including opportunities for micro-teaching, learning about new techniques of teaching and learning, and the like to help faculty become more effective persons and teachers. Teaching evaluations could be used to help identify problems which could be aided by means of a comprehensive faculty development program.

8. The results of teaching evaluation ought to be used, also, to make decisions about retention, promotion, and tenure. It is in the self-interest of the institution, and the entire professoriate, to retain, promote, and award tenure to those persons who are adjudged by the best available evidence to be effective teachers. This is especially true today when we have an abudance of prospective teachers for each

open position; unlike the days of a teacher shortage, there is little justification for rewarding ineffective teaching any more.

9. Most teaching evaluation procedures attempt to learn how well individual teachers are performing within the general university structure. Yet, we know that individuals are severely constrained by their environments; the institutional climate, faculty value scheme, peer group pressures, and institutional organization all impose limitations on the effectiveness of any individual. Further, teaching may be significantly improved by modifying the environment within which a faculty member teaches. Thus, innovations such as cluster colleges, offering alternative educational environments, should be encouraged with vigor at least equal to that propelling teaching evaluation.

10. A few schools have decided that they can best respond to the need to improve instruction by creating teaching resource centers. Although such centers vary in size, structure, and program, they all provide some of the services discussed earlier to help faculty members improve their teaching. Because there will be few additional faculty positions at most schools in the foreseeable future, an increasing need will be to help the existing faculty to grow and develop as teachers. For this reason I think we can and should look forward to these offices becoming the newest entries on the organization charts of many institutions.

It is my conviction that the new directions in research and theory I have suggested will allow us to better understand the complicated dynamics of teachers who make a difference with students and that the suggestions for going beyond the current use of student ratings in practice will allow faculty members to make a greater impact in the education of students.

TABLE 1

STUDENT PERCEPTIONS OF THE WAYS INFLUENTIAL FACULTY MEMBERS HELPED THEM

(In Percentages)
(N = 1127)

STATEMENT	Not at all descriptive (1)	Somewhat descriptive (2)	Quite descriptive (3)	Very descriptive (4)
He or she:				
Was available and open to any discussion	4	17	30	51
Stimulated me intellectually	3	16	35	46
Helped me feel confident of my own abilities	9	18	35	37
Demanded high quality work work from me	11	19	32	37
Interested me in his/her field	10	24	31	35
Encouraged me to inspect my values	31	25	26	17
Advised me about my career plans	31	31	22	16
Made me aware of social issues	36	31	21	13
Counseled me about a personal problem	59	22	9	10
Helped me get a job or scholarship	71	12	8	10

FACULTY PERFORMANCE UNDER STRESS

Mary Jo Clark — Educational Testing Service

Robert T. Blackburn — University of Michigan

The major focus of this research report is on the stresses faculty members feel as they conduct their work and the relationship between these conflicts or pressures and their performance as classroom teachers. The performance measures used in the study are ratings of teaching effectiveness by faculty colleagues and also ratings of teaching effectiveness by students in each professor's classes. Therefore, a second focus will be upon the extent to which student and faculty raters agree about the teaching effectiveness of professors under different conditions of stress and with various personal characteristics.

These data are part of a larger study designed to apply the propositions of role conflict theory and organizational stress to the workings of a small baccalaureate college. The basic notions of this theoretical framework are best presented in diagram form. (See Figure 1.)

Insert Figure 1 about here

The conceptual framework for the study comes from work on role sets and role conflict by Robert Kahn and colleagues (1964) in relation to studies of personal health in organizations. In their theoretical model, both personal characteristics and the organizational environment directly affect outcome variables (e.g., performance on the job, or satisfaction). Additionally, an interaction between the individual and the organization takes place as the person works in the job environment.

This fit between the person and the organization, a created psychological environment, also directly affects outcomes.

Stresses or conflicts in this situation can take many forms, but one of the most common reactions to heavy or ambiguous demands of the job is to feel unduly pressured and loaded down. The psychological environment, or the fit of the person and the organization, will moderate this reaction; some people respond to heavy work demands more quickly or more negatively than others. But, in general, when a focal person says he feels highly overloaded, it is like saying that he feels the pressures are beyond his particular inclination or capacity to cope with them effectively. The central hypothesis of this study is that a person's responses to the stress of role overload will be detrimental to role performance, and that the extent of this effect will be moderated by the enduring personal properties of the person.

Two forms of role overload are selected for primary attention. The first is quantitative (QT) overload, or the discrepancy the individual feels between job requirements and the time available to accomplish them. With professionals, such as faculty members, this time factor is concerned with preferred use of time as well as with the actual number of hours available. The other factor is qualitative (QL) overload, the discrepancy between the demands of the job and the person's sense of being able to meet the demands irrespective of time.

Both quantitative and qualitative overload are expected to lead to impaired job performance, although through somewhat different mechanisms. Quantitative overload, by definition, means the person feels he cannot perform his job in the way expected by all of his role senders because there is too much work for him to do in the time available.

Therefore, their evaluations of his performance are likely to suffer. But if work demands conflict with self-attributed lack of ability or skill, leading to qualitative overload, the effect may be most apparent in a lowered level of job attention and satisfaction. These conditions in turn, may contribute to lower evaluations by others.

A high level of experienced overload in one area can reasonably be expected to increase the level of felt pressure in the other area. For instance, concern about one's ability to perform the work (contributing to high qualitative overload) probably increases susceptibility to feelings of pressure from lack of time (quantitative overload) and may lead to substandard performance. Or too much work to do (high quantitative overload) might contribute to concern about succeeding professionally which would be reflected in feelings of high qualitative overload. Thus, though quantitative overload and qualitative overload are conceptually distinct, they are related, and a high level of either one is expected to affect role performance. A low positive correlation between measurements of quantitative and qualitative overload is expected, and both are expected to correlate negatively with job satisfactions and with independent ratings of job performance.

In addition to the direct effects of work performance diagrammed in Figure 1, this research specifically hypothesized that traits of the person will moderate the relationship between stress and performance. In terms to Figure 1, this hypothesis states that enduring personal characteristics such as level of emotional sensitivity or tendency toward sociability (arrow 1) interact with the conflicts and stresses experienced in the work situation (arrow 5) to demonstrate relationships with work performance that are different from the direct effects of either set of variables considered separately. Examples of predictions from this conception are that high stress will be most damaging to the work performance of faculty members who have a high level of emotional sensitivity, or for those professors who tend toward social independence rather than sociability. This hypothesis proposes that consideration of personal factors along with level of stress will improve our understanding of the relationship between stress and performance.

Subjects

Subjects for this study were faculty members at a small liberal arts college that we will call "Midwest College." Forty-five professors, or 85 percent of all full-time faculty members, provided full information and are included in these results. They represent a variety of fields, backgrounds, and levels of academic experience. The principal faculty roles are teaching and participating in the general activities and operation of the college. Students are average in ability and variety of interests. In these respects, the college is similar to many general-purpose baccalaureate programs across the country. It is neither highly selective nor self-consciously open-door, but middle-of-the-road and, at the time these data were collected, relatively traditional in its view of the teaching-learning process.

Specifically, each faculty member rated every other teacher in his curriculum division on a five-point scale of "teaching effectiveness." Raters were told to "consider those qualities which are important in the evaluation of the skills and practices and products of a classroom teacher, regardless of rank or experience or training of the person being rated."[1]

Student evaluations of teaching effectiveness were obtained from a standard 14 item five-point scale questionnaire the college systematically employed to evaluate all courses each semester. Responses to the question "How would you rate your instructor in teaching effectiveness?" were averaged across all courses taught by a faculty member during the semester in which other data were collected. The professor's mean served as the index of his teaching performance as judged by students.

Faculty members also completed questionnaires on academic attitudes and values, background characteristics, and personal traits. Thirty stress items similar to those used by Mueller (1965) were factor analyzed and yielded results consistent with the factors he obtained from responses by faculty members in a large, research oriented university. The quantitative (QT) overload index was constructed by totaling weighted individual responses to the five items[2] that loaded highest on the factor assigned this label.

---

[1] The method is one of using experts, in this case professional colleagues, to make judgments about quality. Perhaps the best documented recent use of this technique, at least in higher education, is the ACE ratings of doctoral programs (Cartter, 1966; Roose and Anderson, 1971). See Clark & Blackburn (1973) for details concerning the analyses carried out to establish the reliability and validity of the measures used in the study here reported.

[2] Overwhelming workload. Too many things to be done.
The feeling of never having any time.
Not being able to allocate my time and resources as I wish to.
Not enough time to think and contemplate.

The four items[3] loading highest on the factor labeled qualitative (QL) overload were totaled to form an index of this variable. No item included in one index loaded above .25 on the other factor, and most had alternate loadings near zero.

Wherever possible, established measures were used to represent the personal attributes under study. The measures of emotional sensitivity or anxiety (Ax) is the total score on two subscales (22 items) of the I.P.A.T. anxiety scale (Cattell, 1956). The flexibility (Fx) index is the total score from 22 items comprising the flexibility scale on the California Personality Inventory (Gough, 1957). Items used to construct the Self-Esteem (SE) index come from two shorter measures, one by Rosenberg (1965) and the other by Cobb _et al_ (1966). Sociability (So) is defined by Bass's (1967) social interaction scale in the Orientation Inventory. Research Orientation (Res) is more a value than a personality trait and probably is less enduring and stable. This index is a factor score over 22 items concerned with the profession of college teaching, the relative weight assigned to research and teaching as academic role obligations, and preferred teaching styles. The items loading highest on the factor are listed in footnote 4.

---

[3] The desire to succeed.
Not measuring up to the demands of the job, lack of training or knowledge or talent.
Responsibility for and control of people's futures.
Competition to keep up with my colleagues.

[4] Research is the academic man's most important activity.
For me, research obligations are relatively unimportant in contrast to teaching obligations.
It is important for a faculty member to engage in both teaching and research; neither should be stressed in preference to the other.

Questionnaire responses and institutional records also gave data on actual and preferred distribution of work time on differentiated activities, intrinsic and extrinsic job satisfaction, teaching load, committee assignments, and the like as well as providing standard demographic data.

Results

For many of the analyses the respondents were divided as evenly as possible into high and low groups on each personal attribute and on the relevant index of experienced stress. The "high and "low" designations are relative terms and may or may not have any "absolute meaning. For example, this faculty reports an average work week of more than 56 hours. Hence those in the "low" group are still carrying a heavy load. Similarly, on the emotional sensitivity scale (Ax), the total scores of the respondents range from 35 to 68 on a scale running from 22 to 110. The group designated more anxious or emotionally excitable, then has a mean score well below levels associated with serious emotional distress. In the opposite direction, self-esteem scores range from 31 to 53 out of a possible 11 to 55. Thus, in fact, members of the "low" self-esteem group think rather well of themselves. Once more, a high and low are relative terms used only to represent the direction of certain factors in the data analysis.

The stress measure of quantitative overload, representing a discrepancy between time demands and individual preferences for time allocation, demonstrated negative but very low (statistically non-significant) correlations with age, years of experience, rank, and salary. There is also little apparent association between this subjective measure of quantitative overload and available indicators of objective workload, such as teaching load or hours worked per week. This is perhaps not too surprising when we note that almost all faculty members, even division chairmen, teach 10 to 15 hours per week, serve on two to five committees, and average 56 hours per week on the job. The basic work situation is heavy for them all. Instead, the subjective measure of quantitative overload seems to reflect conflict between the individual and the work situation rather than a direct representation of objective workload. For instance, professors with high QT overload scores also say that they feel a lot of pressure from college assignments, regulations, and requests for services.

Intercorrelations of the stress factors and the personal attribute measures are presented in Table 1.

Insert Table 1 about here

As predicted, there is a moderate positive relationship between QT and QL overload (r = .36).[5] The measure of emotional sensitivity (Ax) also shows moderate relationships with stress from time pressure (QT), level of flexibility, and level of self-esteem. In general, however, the intercorrelations of these self-report variables are low, suggesting reasonable independence in measurement as well as conception.

Our first hypothesis derived from the conceptual model stated that high stress (high QT or QL overload) would negatively affect work performance, or rated teaching effectiveness. Figure 2 diagrams mean performance ratings by students and by faculty peers when faculty members are divided into low and high groups on QT and QL overload. Though student ratings are somewhat lower for faculty members who report that they feel a lot of pressure and conflict, the differences are not statistically significant and we must reject the hypothesis. If we consider only experienced stress, there seems to be little effect on the teacher's work performance.

---

[5] Mueller (1965) obtained a correlation of .34 between QT and QL in his study of university professors.

Insert Figure 2 about here

Our last results concern the interaction of enduring personal traits and experienced conflict on performance. For these analyses, faculty members were divided high and low on each stress variable and high and low on each personal characteristic. Mean rated teaching effectiveness as rated by students and by faculty peers were calculated for each of the four cells. Figure 3 diagrams mean performance scores for low and high QT overload and low and high classifications on each of the five personal dimensions.

Insert Figure 3 about here

As can be seen in Figure 3, students and faculty have highly similar patterns of assessment concerning faculty teaching effectiveness. That is, there is general agreement on relatively higher or lower ratings as well as on the effects of stress and the moderation of this effect by personal characteristics. This finding is in accord with correlations above .60 between student and faculty assessments of teaching as reported by Maslow and Zimmerman (1956) and Choy (1970).

Critics of rating procedures for measuring teacher performance often question whether faculty members can (or will) discriminate among their colleagues on this dimension, suggesting that the results are likely to look very flat and uninteresting. Inspection of Figure 3 suggests that this is not the case. Though student ratings tend to exhibit slightly greater variability, and therefore reach levels of statistical significance somewhat more often in these data, faculty colleagues show marked and consistent differences in evaluations of teaching among their peers in relation to two separate indexes of stress.

On teaching effectiveness, students are rating faculty members they have observed in the classroom over the course of a semester; faculty members are rating colleagues in the same curricular division with whom they interact in various professional ways, but generally do not directly observe in the classroom. Factors of low and high stress and low and high personal traits enter into the ratings only insofar as they affect the rater's perception of the effectiveness of the faculty member's teaching. Given the independence of the ratings and the personal variables, there is remarkable consistency between faculty members and students across the five personal conditions. Both sets of raters agree that under high quantitative overload, an otherwise high level of teaching effectiveness definitely drops among faculty members who are more emotionally excitable, are more rigid, have a higher self-esteem, are more independent, and have a higher research orientation. But high QT overload has little apparent effect on the initially lower effectiveness of teachers who are calm, are more flexible, have lower self-esteem, are more sociable, and are more orientated toward teaching than research.

The effects of qualitative overload (Figure 4) on teaching effec-

----

Insert Figure 4 about here

----

tiveness are generally similar, though higher ratings on teaching effectiveness under some high stress conditions are apparent. Again, independent ratings by faculty members and students are very much alike. It appears that high overload stress, whether time or ability related, is harmful to the teaching effectiveness of the kinds of faculty members who tend to get the highest teacher ratings under low overload conditions, but is not particularly harmful (and may even be beneficial) to the teaching of those who receive the lower ratings when stress is low.

Though at first these results for the lower rated teachers appear to be contradictory, they are consistent with the notion of involvement or "creative tension" (Pelz, 1967) as a prerequisite to top-level work among independent professionals working in organizational settings. It could be argued that less excitable, more flexible, more sociable, and more teaching-orientated faculty members are adequate as teachers under conditions of low stress, and they continue to perform at about the same level when they are pushed hard, either qualitatively or quantitatively. In fact, they may even do better as they respond to the challenge. However, their counterparts fall apart under high pressure, particularly when it is time pressure, and their teaching suffers. Already maximally involved under conditions of low stress, the additional pressure can only be disruptive. These kinds of teachers get the highest ratings when they are not too pressured. But, with high pressure, they cannot keep up with the demands and their work suffers.

Both students and faculty give highest teacher ratings to faculty members who have a high research orientation, are socially independent, have a high value of self, are comparatively rigid, and who also do not feel much sense of role overload, either quantitative or qualitative. (See Figures 3 and 4.) Apparently these are the people who thrive on pressure, or who are self-sufficient enough to be relatively oblivious to it. For it is exactly these same kinds of people who are rated much lower in their teaching performance when they also express a high level of work overload.

In summary, these data support the association of role overload stress and performance as moderated by personal traits and values. Some kinds of people are bothered by feelings of pressure while others are less affected or even seem to be challenged by the same condition. However, we should note that even though the performance ratings of some people may actually be higher under high stress, the job satisfactions of these people suffer most under these same high stress conditions. Therefore, stress under any condition carries with it some penalty, though some effects will be reflected most directly in the immediate performance of one's job.

Conclusions

The findings have immediate and telling implications for the managing of colleges and universities and for the people who work in them. Faculty recruitment and retention, work assignment and load, the reward structure of recognition, tenure, and promotion, all need to take into account how performance is affected by stress and moderated by personal characteristics.

For example, students rated the more rigid faculty members under low overload stress as their most effective teachers. Colleagues too valued conformity in relation to ratings of teaching effectiveness.

Presumably, when not under particular pressure, the more rigid faculty member is better organized and prepared while also sufficiently relaxed in class to be viewed as a good teacher. But, when things get tight, he tends to get dogmatic and flustered, and his teaching performance deteriorates. How he is viewed on a teacher evaluation form, then, will depend in part on his other work and life circumstances. There are two major implications for interpretation of his ratings: first, a pattern of rating rather than ratings at any one time should be used in any decision-making situation. Second, ratings should be interpreted in the context of other information about the individual.

Three other illustrations point up implications. First, high overload appears to be detrimental to performance among those least able to cope with stress--the excitable, the least flexible, the socially more isolated, and the strongly research oriented. Second, faculty who suffer most under high overload are the individuals least able to deal constructively with frustration and discouragement, the persons for whom increased anxiety from poor evaluations by students and peers (together with heavy work pressures) are apt to be most counter-productive. More rigid and more socially independent faculty are apt to withdraw further into themselves under increasing pressure. For the research oriented, evaluation in teaching and service become increasingly frustrating because they are the areas of least important personal professional concern.

Third, the findings raise questions regarding a growing student practice, making public faculty evaluations of teaching. The student argument is persuasive. As clients they are entitled to full market information. Consumer reports on faculty provide a basis on which students take or do not take courses. The student concern for improving teaching on campus is genuine. So is their belief that publicly identifying weaker teachers will produce improvement. However, their technique assumes all faculty could teach better if they would only try harder and work at it more. Maybe they can, although Hildebrand (1972) has found that the best and worst judged teachers give equal time to the activity. The personality data in this study and the consequences of stress suggest that for some faculty public ratings will have consequences just the opposite from what is desired.

References

Bass, B. M. Social Behavior and the orientation inventory. *Psychological Bulletin*, 1967, 68, 4, 260-292.

Cartter, A. M. *An Assessment of Quality in Graduate Education*. Washington, D. C.: American Council on Education, 1966.

Cattell, R. B. *Handbook:* The I.P.A.T. anxiety scale. Champaign, Illinois: Institute for Personality and Ability Testing, 1956.

Choy, C. The relationship of college teacher effectiveness to conceptual systems orientation and perceptual orientation. Unpublished Ph.D. dissertation, Colorado State College, 1969.

Clark, M. J. A study of organizational stress and professional performance of faculty members in a small four-year college. Ph.D. dissertation in progress, The University of Michigan, 1973.

Clark, M. J., & Blackburn, R. T. Assessing faculty performance: A test of method. submitted for publication, 1973.

Cobb, S., Brooks, G. H., Kasl, S. V., & Connelly, W. E. The health of people changing jobs: A description of a longitudinal study. *American Journal of Public Health*, 1966, 56, 1476-1481.

Gough, H. G. *Manual for the California Psychological Inventory*. Palo Alto, California, Consulting Psychologists Press, Inc., 1957.

Hildebrand, M. How to recommend promotion for a mediocre teacher without actually lying. *Journal of Higher Education*, 1972, 42, 1, 44-62.

Kahn, R. L., Wolfe, D. M., Quinn, R. P., Snoeck, D. J., & Rosenthal, R. A. *Organizational Stress: Studies in Role Conflict and Ambiguity*. New York: John Wiley and Sons, Inc., 1964.

Maslow, A. H., & Zimmerman, W. College teaching ability, activity, and personality. *Journal of Educational Psychology*, 1956, 47, March, 185-189.

Mueller, I. Workload of university professors. Unpublished doctoral dissertation, University of Michigan, 1965.

Pelz, D. C. Creative tensions in the research and development climate. *Science*, 1967, 157, July 14, 160-165.

Roose, K. D., & Anderson, D. J. *A Rating of Graduate Programs*. Washington D. C.: American Council on Education, 1970.

Rosenberg, M. *Society and the Adolescent Self-Image*. Princeton University Press, 1965.

Table 1

Intercorrelations of Stress and Personal Measures

	QT	QL	Ax	Fx	SE	So	Res
QT	--						
QL	.36	--					
Ax	.31	.11	--				
Fx	-.21	-.01	-.45	--			
SE	-.28	-.15	-.36	-.26	--		
So	-.20	.01	.07	.18	-.09	--	
Res	-.29	.01	-.06	-.02	.07	.16	--

BEST COPY AVAILABLE

Figure 1

CONCEPTUAL FRAMEWORK

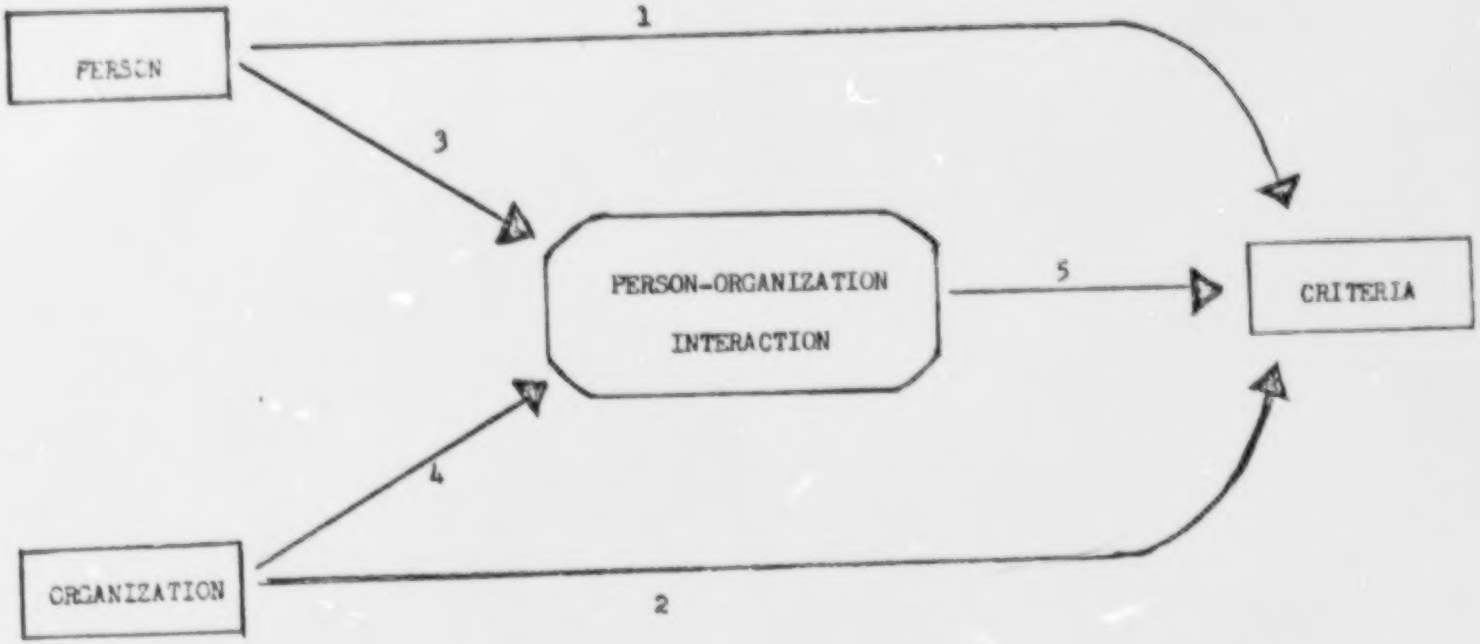
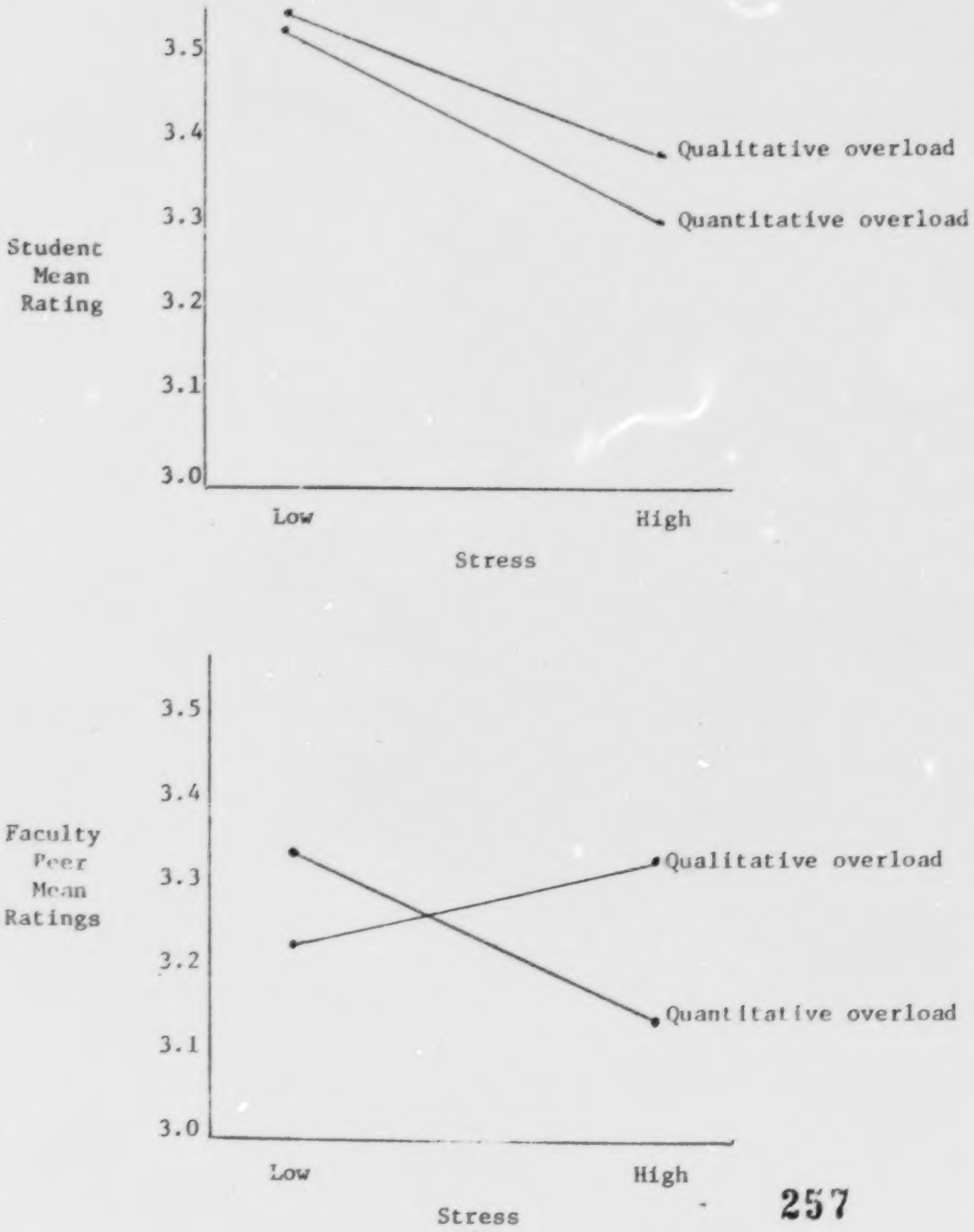


Figure 2

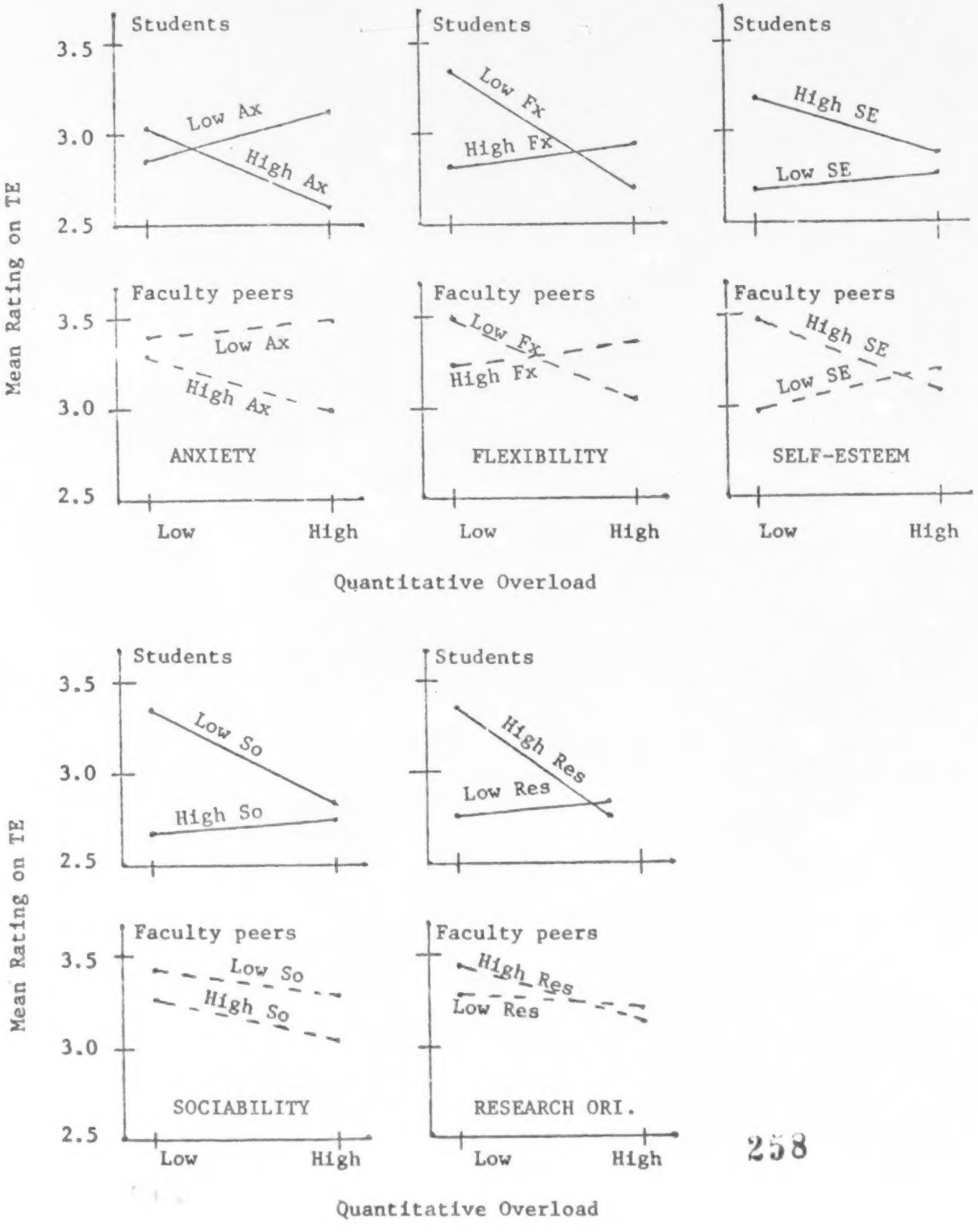
Direct Effect of Stresses
on Student and Faculty Ratings
of Teaching Effectiveness



BEST COPY AVAILABLE

Figure 3

Rated Teaching Effectiveness and Quantitative Overload



BEST COPY AVAILABLE

Figure 4

Rated Teaching Effectiveness and Qualitative Overload